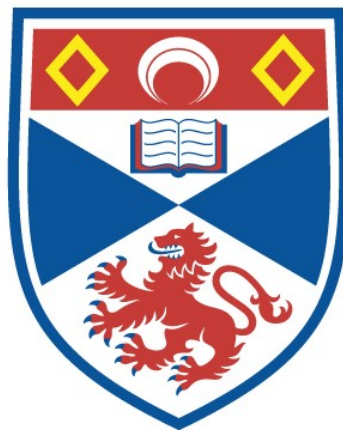


CORPUS LINGUISTICS FOR THE EXPLORATION OF LEGAL PRECEDENT

Evan David Brown

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2019

Full metadata for this thesis is available in
St Andrews Research Repository
at:

<http://research-repository.st-andrews.ac.uk/>

Please use this identifier to cite or link to this thesis:

<http://hdl.handle.net/10023/18188>

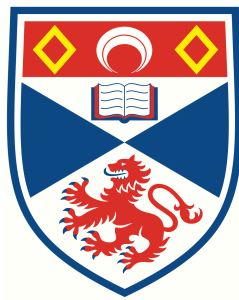
This item is protected by original copyright

This item is licensed under a
Creative Commons License

<https://creativecommons.org/licenses/by-nc-nd/4.0>

Corpus Linguistics for the Exploration of Legal Precedent

Evan David Brown



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of
Doctor of Philosophy
at the University of St Andrews

February 2019

ABSTRACT

A deterioration of legal research skills has become a critical issue for lawyers. This thesis examines the causes of the problem under English law and specifically addresses how current technology for legal research contributes to or ameliorates the skills deficit. England has a “*common law*” system where the state of the law is determined by precedents laid down in previous cases. This linkage between cases means that lawyers have to assimilate a large amount of written language in order to understand the key elements of prior judgements. There are approximately two million criminal cases and another two million civil cases heard in England each year [116], from which the important decisions are reported.

This thesis first analyses the way in which lawyers work in conducting legal research. The findings indicate that the move online has improved access to legal information but it has compromised the ability of practitioners to identify high-quality precedent and to discard information which is not relevant. A side-effect is the marginalisation of trained law librarians and their curation skills which contributes to the problem. Existing platforms prioritise comprehensive data coverage over delivering a curated research environment. A need to better train lawyers in the skills of critical thinking and linguistic analysis through computer-based tools is identified.

This thesis proposes that linguistic analysis techniques from the domain of corpus linguistics can help. Single-context legal research tools which minimise the need to switch between different applications are also required. Effective working is currently compromised by having to navigate between many different tools. The development of effective collaboration skills is of particular importance. Lawyers must work well in teams in order to prepare cases effectively.

The Legal Research and Collaboration platform is the prototype application which results from this research. It is a software system for legal research within teams of lawyers. Experiments establish how effectively LARC works for both practising lawyers and for law students. A foundation for future work is laid because the software is entirely open source and is based upon open access legal data. The results show that critical barriers which result in poor legal research skills can be ameliorated by well-designed computer-based tools.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisors at the University of St Andrews - Professor Aaron Quigley and Dr Miguel Nacenta. The SACHI Human Computer Interaction group, of which I have been a member, is friendly and informal. It is also a highly engaging and productive place to work. My supervisors have supported me throughout my studies whilst always encouraging me to make sure that the quality of my work was as high as possible.

I owe a particular debt of gratitude to my parents. They have always believed in my ability to complete this research. My father, Ken Brown, has read everything that I have produced over the last five years and has provided valuable feedback and suggestions for improvement. His contributions often made me consider the broader context of the research which could easily have been forgotten as I became ever more engrossed in and close to my chosen field of study. My mother, Maureen Brown, has offered encouragement and reassurance particularly when I have encountered practical problems in my work that could have become all-consuming.

I would like to thank my brother, Calum Brown, for his guidance on the practicalities of studying for a doctorate. My frequent visits to stay with him and his family in Germany have allowed me to maintain a broad perspective in which the role of my research work was kept in proportion to leading a rounded life. This has been greatly helped by the reliable supply of tasty German beer which is available whenever I go to stay!

Finally, I would like to dedicate this thesis to my friend, John Sinclair. John was Professor of Modern English Language at the University of Birmingham between 1965 and 2000 and President of the Tuscan Word Centre in Florence after his retirement. In these roles, he was also an eminent first-generation corpus linguist. I knew and worked with John all too briefly before his death in 2007 - but he pointed me in the right direction nonetheless. I hope that my efforts to apply corpus linguistics to the legal domain and to legal texts would meet with his approval.

Evan Brown

Candidate's declaration

I, Evan David Brown, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 80,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in October 2013.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date

Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date

Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Evan David Brown, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date

Signature of candidate

Date

Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Evan David Brown, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date

Signature of candidate

CONTENTS

List of Figures	vii
List of Tables	xi
1 Introduction	1
1.1 <i>An example of the evolution of precedent</i>	2
1.2 <i>Thesis statement</i>	6
1.3 <i>Scope</i>	6
1.4 <i>A note on methodology and the role of BAILII</i>	8
1.5 <i>Contributions</i>	10
1.5.1 <i>The move to open knowledge: open data and open source</i>	10
1.5.2 <i>Main research contributions</i>	13
1.5.3 <i>Additional outcomes</i>	16
1.6 <i>Thesis outline</i>	16
1.6.1 <i>Research question</i>	18
I BACKGROUND & RELATED WORK	21
2 Background: Law	23
2.1 <i>Introduction</i>	23
2.2 <i>The development of English common law</i>	24
2.3 <i>The rule of law</i>	26
2.4 <i>Sources of law: legislation and its judicial interpretation</i>	28
2.5 <i>Sources of law: the doctrine of precedent or stare decisis</i>	31
2.6 <i>Research method: The focus on English law</i>	35
2.7 <i>The diminishing role of domestic precedent</i>	36
2.8 <i>Language and the law</i>	39
2.9 <i>The nature of legal training</i>	45
2.10 <i>Legal training: Simulation and mootng</i>	48
2.11 <i>A boundary limitation: Other forms of applied training for lawyers</i>	55
2.12 <i>Conclusion</i>	56

3	Related Work	59
3.1	<i>Introduction</i>	59
3.2	<i>Computers and the law - information discovery platforms</i>	60
3.3	<i>How lawyers work with technology</i>	65
3.4	<i>The research skills deficit</i>	70
3.5	<i>The growing role of corpus linguistics</i>	73
3.6	<i>Corpus interfaces: KWIC and other visualisations</i>	79
3.6.1	<i>Alternatives and additions to the KWIC paradigm</i>	81
3.7	<i>Visualising the law</i>	82
3.8	<i>Collaboration environments for lawyers</i>	85
3.9	<i>Conclusion</i>	88
II	SETTING THE SCENE	91
4	Legal Research	93
4.1	<i>Thesis process</i>	93
4.2	<i>Introduction</i>	93
4.3	<i>Methodology</i>	94
4.3.1	<i>Contextual inquiry</i>	94
4.3.2	<i>Solicitor interviews</i>	95
4.3.3	<i>Lawyer survey</i>	96
4.4	<i>Analysis procedure</i>	96
4.4.1	<i>Contextual inquiry</i>	96
4.4.2	<i>Solicitor interviews</i>	98
4.4.3	<i>Lawyer survey</i>	98
4.5	<i>Results</i>	98
4.5.1	<i>Tasks and activities</i>	99
4.5.2	<i>Collaboration</i>	110
4.5.3	<i>Tools</i>	117
4.6	<i>Discussion</i>	126
4.6.1	<i>Modes and Tools of Collaboration and Communication</i>	127
4.6.2	<i>Note taking</i>	128
4.6.3	<i>Search and Supervision</i>	129
4.6.4	<i>Opportunities and Recommendations</i>	130
4.7	<i>Conclusion</i>	132
5	LARC - The LegAI Research and Collaboration platform	133
5.1	<i>Thesis process</i>	133
5.2	<i>Introduction</i>	134

5.3	<i>Current barriers to effective working practice</i>	134
5.4	<i>The overall system architecture</i>	137
5.5	<i>Using open access legal data</i>	139
5.5.1	<i>The limited extent of open access legal data under English law</i>	139
5.5.2	<i>The web scraper</i>	141
5.6	<i>Database design and structure</i>	143
5.6.1	<i>How to identify relevant content</i>	145
5.7	<i>Enabling search facets</i>	147
5.8	<i>Identifying case connections and visualising them</i>	155
5.9	<i>Supporting note taking</i>	157
5.10	<i>Instant messaging and chat facilities</i>	158
5.11	<i>Auditing tools</i>	159
5.11.1	<i>Saved contexts</i>	162
5.11.2	<i>Statistics</i>	163
5.11.3	<i>Sessions and history</i>	166
5.11.4	<i>Interesting phrases</i>	166
5.12	<i>Summary: Bringing it all together - the complete LARC user interface</i>	174
5.13	<i>Key evaluations</i>	175
5.13.1	<i>Differences between LARC and JustCite</i>	175
5.13.2	<i>Citation Accuracy</i>	178
5.13.3	<i>Treatment Accuracy</i>	178
5.13.4	<i>Keyword Accuracy</i>	180
5.13.5	<i>Differences in analysis between JustCite and LARC: Fairchild v Glenhaven</i>	180
5.14	<i>Discussion</i>	181
5.14.1	<i>Mapping design decisions to barriers</i>	181
5.14.2	<i>Limitations and opportunities</i>	183
5.15	<i>Conclusion</i>	184
III	EXPLORING LEGAL LANGUAGE	185
6	Integrating Language Search	187
6.1	<i>Thesis process</i>	187
6.2	<i>Introduction</i>	187
6.3	<i>Designing a corpus for use by lawyers</i>	193
6.3.1	<i>Why use a corpus linguistics approach?</i>	193
6.3.2	<i>The need for new legal corpora</i>	194
6.3.3	<i>The argument for clean text</i>	196
6.3.4	<i>The problem of structure</i>	196

6.3.5	<i>Preparing the LARC corpora</i>	198
6.4	<i>Moving away from keywords to collocations</i>	200
6.4.1	<i>Extracting collocations from the corpus</i>	203
6.4.2	<i>Search as a workflow - drill down</i>	208
6.5	<i>From collocations to concordances</i>	209
6.6	<i>Looking at variation outside the node</i>	212
6.7	<i>An infrastructure for caching results</i>	215
6.8	<i>Tying it all together - the complete interface</i>	217
6.9	<i>Algorithmic evaluation</i>	218
6.9.1	<i>Collocates of negligence</i>	219
6.9.2	<i>Typical concordance for hot</i>	221
6.10	<i>Conclusion</i>	224
IV	EVALUATION AND DISCUSSION OF RESULTS	227
7	System Evaluation	229
7.1	<i>Thesis process</i>	229
7.2	<i>Introduction</i>	229
7.3	<i>Identifying the evaluators</i>	230
7.4	<i>Evaluation task</i>	232
7.5	<i>Metrics used</i>	234
7.6	<i>Data collection approach</i>	234
7.7	<i>Results</i>	240
7.7.1	<i>Quantitative metrics - SUS ratings for LARC</i>	240
7.7.2	<i>Qualitative measures</i>	241
7.7.3	<i>Critical incidents</i>	246
7.7.4	<i>Priorities for fixing incidents and future work</i>	250
7.8	<i>Conclusion</i>	251
8	Discussion	253
8.1	<i>Introduction</i>	253
8.2	<i>Resolving the research question</i>	253
8.2.1	<i>The results of P1 - How lawyers conduct legal research</i>	254
8.2.2	<i>The results of P2 - Technology and working practice</i>	255
8.2.3	<i>The results of P3 - Integrating open source software</i>	256
8.2.4	<i>The results of P4 - Supporting collaboration</i>	257
8.2.5	<i>The results of P5 - Designing search interfaces for lawyers</i>	258
8.2.6	<i>The results of P6 and P7 - LARC's usability and feedback for future work</i>	259

8.2.7	<i>Answering the main research question - How to apply corpus linguistics to legal research</i>	259
8.3	<i>Conclusion</i>	261
9	<i>Conclusion</i>	263
	<i>References</i>	269
V	APPENDICES	287
	Appendix A Setting the scene: contextual inquiry, interviews and survey	289
A.1	<i>Mooting Work Roles</i>	290
A.2	<i>Mooting Spatial Arrangement</i>	292
A.3	<i>Hierarchical Task Analysis</i>	293
A.3.1	<i>Reading the problem question</i>	293
A.3.2	<i>Identifying seed cases</i>	294
A.3.3	<i>Splitting the problem question</i>	295
A.3.4	<i>Searching for relevant cases</i>	296
A.3.5	<i>Searching for relevant legislation</i>	297
A.3.6	<i>Identifying relevant journal articles</i>	298
A.3.7	<i>Building an argument</i>	299
A.3.8	<i>Identifying the counter-argument</i>	300
A.3.9	<i>Preparing a speech</i>	301
A.3.10	<i>Preparing a rebuttal</i>	302
A.4	<i>Flow diagrams</i>	303
A.4.1	<i>Reading the problem question</i>	303
A.4.2	<i>Identifying seed cases</i>	304
A.4.3	<i>Splitting the problem question</i>	305
A.4.4	<i>Searching for relevant cases</i>	306
A.4.5	<i>Searching for relevant legislation</i>	307
A.4.6	<i>Identifying relevant journal articles</i>	308
A.4.7	<i>Building an argument</i>	309
A.4.8	<i>Identifying the counter-argument</i>	310
A.4.9	<i>Preparing a speech</i>	311
A.4.10	<i>Preparing a rebuttal</i>	312
A.5	<i>Artefact model</i>	313
A.6	<i>Social model</i>	314
A.7	<i>Affinity diagrams</i>	315
A.7.1	<i>Affinity diagram A</i>	315
A.7.2	<i>Affinity diagram B</i>	316

CONTENTS

A.7.3	<i>Affinity diagram C</i>	317
A.7.4	<i>Consolidated affinity diagram</i>	318
A.8	<i>Barrier severity matrix</i>	319
A.9	<i>Mooting interview questions</i>	320
A.10	<i>Solicitor interview questions</i>	324
A.11	<i>Lawyer survey questions</i>	328
A.12	<i>Work role data from solicitor interviews</i>	329
A.13	<i>Drivers of collaboration from lawyer survey</i>	331
Appendix B	<i>LARC - The LegAl Research and Collaboration platform</i>	335
B.1	<i>Final refined wireframe sketches for LARC</i>	336
B.2	<i>Wireframe details for LARC interface elements</i>	348
Appendix C	<i>Evaluation products</i>	353
C.1	<i>The evaluation questionnaire</i>	353

LIST OF FIGURES

1.1	Overview of the structure of this thesis, the main topic areas concerned and key contributions.	14
1.2	Overview of the UX lifecycle as presented by Hartson and Pyla in [71]	18
4.1	Work roles diagram from the interviews with solicitors. Discontinuous lines indicate typical connections between the type of role and the professional title. Assistants that are not solicitors or barristers/advocates can do any of the roles and are not shown in the figure (for clarity). . .	101
4.2	The activity transition chord diagram. The nodes around the outside of the circle are the low-level generic tasks from Table 4.7. The size of each node shows the total number of transitions to that state that we recorded. The edges are the number of transitions from a source state to a destination state. The colour of each edge shows the source state that the transitions come from. Loops within the same state are shown as edges that originate and terminate in the same state.	107
4.3	The work roles model from mooting observation	108
4.4	The CSCW Matrix for legal case preparation.	113
4.5	The physical arrangement of mooters across the two sessions.	114
4.6	The collaboration diagram. There is an activity bar for each person which shows which of the low-level activities the participant was engaged in. The key for the colours is given in the legend and corresponds to the low-level tasks described in Table 4.7. A black line along the bottom of the activity bar for a participant shows where they talked to their partner. A green bar indicates when they glanced at their partner's laptop screen. Activity bars cross where real-time collaboration took place.	116
5.1	The LARC infrastructure diagram.	137
5.2	The Entity Relationship Diagram for the LARC database	143
5.3	The initial search interface in LARC with case facet options displayed.	150
5.4	The case name auto-suggestion system in the LARC search interface .	151
5.5	The date search facet in the LARC interface	152
5.6	The judge search facet - demonstrating the waterfall - in the LARC interface.	153
5.7	The sentiment trainer user interface	154
5.8	Citation visualisation.	155
5.9	Integrated Timeline Visualisation	159
5.10	The saved context view in the LARC interface.	161

LIST OF FIGURES

5.11	Document statistics for the current pad in LARC.	164
5.12	The sessions manager for the shopping cart in LARC.	165
5.13	The LARC Research View Interface.	168
5.14	The case-centric search interface in JustCite.	169
5.15	The case reading interface in JustCite.	170
5.16	The cited cases view in JustCite.	171
5.17	The key passages view in JustCite.	172
5.18	The precedent diagram in JustCite.	173
6.1	The standard method of language search presentation in Ravel Law. . .	189
6.2	The corpus-based approach to presenting language search results. . . .	190
6.3	The key sources dialogue which is extracted from the corpus in LARC. .	191
6.4	The concordance view for a query on <i>negligence</i> in LARC.	192
6.5	A simple CQL query for <i>contributory negligence</i>	198
6.6	Sample vertical text output from the corpus preparation program. . . .	199
6.7	The observed and expected ratio equation for collocate significance in LARC	203
6.8	The z-score algorithm for collocate significance in LARC	205
6.9	The t-score algorithm for collocate significance in LARC	205
6.10	The mutual information algorithm for collocate significance in LARC .	206
6.11	The log-likelihood algorithm for collocate significance in LARC	207
6.12	The Typical concordance line significance algorithm in LARC.	211
6.13	The tree concordance for a query on <i>battery</i>	214
6.14	The fuller co-text for a search hit on <i>negligence</i> from the returned concordance.	215
6.15	The RabbitMQ federation manager used to precompute frequent queries in LARC.	217
6.16	A sample Typical concordance for the query <i>hot</i>	222
6.17	A sample filtered Typical concordance for the query <i>hot</i>	223
7.1	Scores from the evaluation participants on SUS questions relating to the utility of the LARC prototype. The scoring scheme is a five point Likert scale from 1 (strongly disagree) to 5 (strongly agree). See Section 7.7.1 for an explanation of the diagram.	236
7.2	Scores from the evaluation participants on SUS questions relating to the ease of use of the LARC prototype. The scoring scheme is a five point Likert scale from 1 (strongly disagree) to 5 (strongly agree). See Section 7.7.1 for an explanation of the diagram.	237
7.3	Scores from the evaluation participants on SUS questions relating to the quality of integration and ease of adoption of the LARC prototype. The scoring scheme is a five point Likert scale from 1 (strongly disagree) to 5 (strongly agree). See Section 7.7.1 for an explanation of the diagram.	238
7.4	Scores from the evaluation participants on SUS questions relating to the learning curve of and user confidence in using the LARC prototype. The scoring scheme is a five point Likert scale from 1 (strongly disagree) to 5 (strongly agree). See Section 7.7.1 for an explanation of the diagram.	239
7.5	The evaluation session recording and playback interface.	246

A1	Work roles identified from video analysis of the mooting exercises . . .	290
A2	Work roles identified from audio analysis of the solicitor interviews . .	291
A3	The physical arrangement of the mooting preparation group	292
A4	HTA for reading the problem question	293
A5	HTA for identifying seed cases	294
A6	HTA for splitting the problem question	295
A7	HTA for searching for relevant cases	296
A8	HTA for searching for relevant legislation	297
A9	HTA for identifying relevant journal articles	298
A10	HTA for building an argument	299
A11	HTA for identifying the counter argument	300
A12	HTA for preparing a speech	301
A13	HTA for preparing a rebuttal	302
A14	Flow diagram for reading the problem question	303
A15	Flow diagram for identifying seed cases	304
A16	Flow diagram for splitting the problem question	305
A17	Flow diagram for searching for relevant cases	306
A18	Flow diagram for searching for relevant legislation	307
A19	Flow diagram for identifying relevant journal articles	308
A20	Flow diagram for building an argument	309
A21	Flow diagram for identifying the counter-argument	310
A22	Flow diagram for preparing a speech	311
A23	Flow diagram for preparing a rebuttal	312
A24	Artefact diagram between work roles	313
A25	Social diagram showing interactions between mooting actors	314
A26	Affinity diagram A	315
A27	Affinity diagram B	316
A28	Affinity diagram C	317
A29	Consolidated affinity diagram from the three initial iterations	318
A30	Severity matrix for identified barriers to effective working practice . .	319
B1	LARC Interface: Wireframes (set 1)	336
B2	LARC Interface: Wireframes (set 2)	337
B3	LARC Interface: Wireframes (set 3)	338
B4	LARC Interface: Wireframes (set 4)	339
B5	LARC Interface: Wireframes (set 5)	340
B6	LARC Interface: Wireframes (set 6)	341
B7	LARC Interface: Wireframes (set 7)	342
B8	LARC Interface: Wireframes (set 8)	343
B9	LARC Interface: Wireframes (set 9)	344
B10	LARC Interface: Wireframes (set 10)	345
B11	LARC Interface: Wireframes (set 11)	346
B12	LARC Interface: Wireframes (set 12)	347
B13	LARC Interface: Citation layout wireframe	348
B14	LARC Interface: On demand citation node loading	349
B15	LARC Interface: Potential substrate wireframe	350

LIST OF FIGURES

B16 LARC Interface: Substrate elaboration wireframe	351
---	-----

LIST OF TABLES

1.2	The process that will be followed in order to answer the research question that is posed in this thesis.	19
4.1	The separate studies which provide data that is analysed in this chapter, together with the scope of each study.	98
4.2	Percentage time spent on preparing for litigation by job role.	99
4.3	Sub-activities of preparing for litigation, in order of frequency, from the survey data.	100
4.4	Low-level taxonomy of tasks and topics in the legal profession from interviews	104
4.5	High level task descriptions from mooting observation.	105
4.6	Low level activities and sub-activities derived from mooting observation	105
4.7	Task timings from mooting observation. The colours represent encodings of activities in Figure 4.6	106
4.8	Work roles and sub-roles identified in the mooting observation	110
4.9	High level task ontology of collaboration tasks.	111
4.10	Percentage of working time spent collaborating by legal job description	112
4.11	Ranked order of tool use by frequency of response (derived from survey)	117
4.12	Notetaking split between physical and digital (from survey)	118
4.13	Average percentage of time spent taking notes by job title (from survey)	119
4.14	Tools that drive collaboration in the profession.	119
4.15	Satisfaction levels by job description (level of seniority). Standard deviation in these results was 24.2%.	120
5.1	Key barriers to effective working practice from the contextual inquiry, lawyer interviews and online survey which are reported in Chapter 4.	135
5.2	Comparison of citation coverage in ten key cases: JustCite and LARC .	179
5.3	Comparison of treatment accuracy in ten key cases: JustCite and LARC	179
5.4	Comparison of keyword category labels in ten key cases: JustCite and LARC	180
6.1	The contingency table for log-likelihood calculations.	206
6.2	The top twenty collocates for <i>negligence</i> under the observed/expected (1), Z-score (2), T-score (3), Mutual Information (4) and Log Likelihood (5) algorithms.	220
7.1	Initial breakdown of evaluator interest by job role	231
7.2	Final breakdown of evaluators who completed the exercise by job role	232
7.3	Overall SUS scores for the LARC system	240

LIST OF TABLES

7.4	Proposed fix priority of critical incident reports - on a scale of 1 - immediate fix to 3 - desirable feature	251
8.1	A consolidated account of the process that has been followed in this thesis in order to answer the main research question.	254
A1	Work roles data extrapolated from discussion in the solicitor interviews.	330

INTRODUCTION

Research over two decades in the United States ([52], [127]) has highlighted widespread concern amongst legal practitioners that the specialist research skills of students and new entrants to the legal profession are becoming poorer. The original motivation for this thesis was to investigate the problem; to ascertain whether it existed in England and Wales and to find out how technology can be used to ameliorate or reverse skill deficits. The central idea here is that sensible attempts at software integration and design in accordance with principles of usability engineering and user-directed development can help to make technology for legal research better suited to solving skills deficit issues. With such a focus on software usability and human-centred design principles, this thesis initially presents a novel study to determine whether declining research skills are seen to effect English lawyers. It then establishes how technology fits into the picture, both in terms of the already evident capabilities in software to solve problems for lawyers and also in terms of problems which current software and technical approaches themselves introduce to the legal domain.

It is argued that modern online research tools for lawyers prioritise information coverage over effectiveness of access. The intrusion of general Internet search norms in an environment where ease of access to legal information has marginalised the roles of trained law librarians is key. Librarians and subject experts who understand and can impart a systematic appreciation to students of the quality of different sources and the levels of legal precedent have critical roles in high quality legal education. Legal precedent means that, in common law systems like England and Wales, the state of the law on any principle is governed by binding judicial decisions from different senior levels of the court hierarchy

in previous cases. Precedent evolves and expands over time as further cases on similar facts are presented. The hierarchy of the courts means that judgements from different levels of the judicial system carry differing weights of significance and can be binding on lower courts in new situations to differing extents.

This thesis suggests that the only way to properly understand and apply existing precedent to new facts is to analyse and to closely understand the language used by judges in their rationales for arriving at previous decisions. The meaning that can be ascribed to judicial language is constrained by previous language and linguistic constructions which have been developed in earlier cases. Language in general, and legal language in particular, does not operate within an *open choice* descriptive model. It relies upon idioms, metaphor and formulations that have been set out before which are built upon, changed, expanded or invalidated in new situations.

1.1 An example of the evolution of precedent

The principle that you must not injure your neighbour was enshrined in law as part of the modern tort of negligence in *Donoghue v Stevenson* [1932] AC 562. This was achieved through the language of Lord Atkin - that “*you must take **reasonable care** to avoid **acts or omissions** which you can **reasonably foresee** would be likely to injure **your neighbour***”. The facts in *Donoghue* were to do with a bottle of ginger beer which was purchased by the plaintiff. It contained a decomposing snail which she only discovered after drinking some of the liquid.

Later, in *Hedley Byrne & Co Ltd v Heller & Partners Ltd* [1964] AC 465, the scope of a *duty of care* was expanded to cover new facts. The issue under consideration was whether a negligent statement made by a bank could give rise to a breach of a duty of care when it was relied upon by the defendant and subsequently caused them economic loss. The court overturned previous authorities in ruling that a negligent statement could result in a breach of a duty of care where it could be shown that “*a **special relationship** existed between the parties.*” A “*special relationship*” was defined in the following language:

*“...where it is plain that the party seeking the information or advice was trusting the other to exercise such a **degree of care** as the circumstances required, where it was **reasonable** for him to do that, and where the other*

*gave the information or advice when he **knew or ought to have known** that the enquirer was relying on him."*

This idea of a "*special relationship*" was further defined in *Caparo Industries plc v Dickman* [1990] 2 AC 605, where Lord Devlin said that the term meant "*a bond of **close proximity** broadly equivalent to a **contractual relationship***". So the law on duty of care has developed since 1932, establishing new duties in an incremental fashion to fit different facts whilst preserving the idea of negligence as a tenet of tort law. The word groups highlighted bold in the above sentences from the different judgements indicate areas of conceptual complexity which need to be defined by developing an understanding of judicial language in both previous and subsequent cases. The individual terms may not themselves be defined in the judgements where they arise. An important part of the job of a lawyer is to be able to define and understand these constructions by careful and structured reference to other linked sources.

It is the development of this linguistic facility and the ability to understand nuance and language choice which must be encouraged in order to allow for effective legal research skills. It is suggested that a preoccupation at present with broad-based semantic search techniques and artificial intelligence in the law could result in opaque and complex machine learning algorithms being applied to legal information in order to present lawyers with ready-made interpretations of what is considered to be important information. There is a substantial risk that this will be achieved without communicating a corresponding appreciation of how or why cases and the language within them are deemed to be significant.

Semantic search considers the context of keyword hits but it relies too often on structured data, pre-determined ontologies and processing techniques like entity recognition which are subjective, purposive and expensive to facilitate. Witness the recent establishment of research projects to encourage and develop explainability in machine learning and artificial intelligence algorithms [50]. These come amid fears that automatic decision making is becoming too complex to be understandable to system users.

The thrust of this research diverges from artificial intelligence approaches. It is suggested that lawyers need to be furnished with an understanding of linguistic significance which is transparent, which trains them to make their own decisions about weight and which enables them to develop well-grounded facilities for

critical thinking and for determining the importance of precedent. Artificial intelligence is a poor term because human intelligence is not analogous to machine learning or to other techniques of computational pattern recognition and decision making. These technologies have a place in training lawyers but they must be focused towards developing expertise in the lawyers themselves. The challenge is to ameliorate problems of information overload so that the human user becomes able to evaluate the quality of legal information for themselves.

The adversarial environment of a courtroom is concerned, to a large extent, with arguments about the ambiguities and affordances of specific linguistic constructions in previous cases and their suitability or otherwise for application on a new and specific set of facts. This is a task for a well-qualified, expert lawyer - not for an automated “expert” system. The acquisition of such a facility is threatened by an over-reliance on algorithms which make key decisions for the user without properly telling them why a particular conclusion has been reached.

It is proposed that techniques for language selection, filtration and presentation from corpus linguistics - a domain which concerns the study of large collections of written language using computers - can help to develop legal research skills. These techniques are based on various measurements of frequency and word co-occurrence in text which can be fairly transparent to the user. The objective is to allow searches of case law and legislation which move beyond isolated keyword identification or predetermined navigation based on entity identification and ontology mapping to an analysis of units of meaning within unstructured text. This thesis suggests that the lowest indivisible linguistic unit of meaning is the collocation, or a pair of words which occur in close proximity to one another within written and spoken text more often than chance itself would dictate.

The theory arises from the highly influential linguistic philosophies of Ludwig Wittgenstein and John Sinclair which, despite their differences, point out that words do not have atomic meanings but that they derive their senses from backgrounds of usage; a principle often expressed as “*meaning is use*” [194] or through Sinclair’s admonition to “*trust the text*” [164]. This is particularly true in a legal context thanks to the doctrine of precedent. People do not select words one at a time in order to make themselves understood. We talk in idioms and in metaphor, the construction of which involves selecting predetermined building blocks of multiple words which are strongly associated together.

Therefore keyword matching and narrow contextual classification in search tools is fundamentally unsuitable for understanding the thrust and underlying principles behind judicial findings. There is a need to allow lawyers to decide what precise linguistic meaning they are interested in through an iterative navigational search interface which presents search results as they appear in a representative database of case law and legislation (the corpus).

This research presents novel ways to integrate collocation and phrase-based searching into a tool for developing legal research skills in law students and early-stage practitioners. The findings indicate that legal research is often conducted in groups, both in simulated mooted activities at university and within the profession by lawyers of different levels and specialisms. Thus the proposed tool is an integrated, single-context application for enabling collaborative legal research and linguistic analysis with computers. This software is called *LARC*, or the *Legal Research and Collaboration* platform. The accessibility of standard presentational paradigms from corpus linguistics, which are designed to be intelligible to trained linguists, is problematic.

The dominant *KeyWord In Context* presentation is particularly poor for fostering an understanding of language variation and nuance in collocated words outside the search hit (or node) in expert users who are not linguists. It is suggested that an appreciation of this variation is key in legal education because students and practitioners need to understand the spread of how terms are used so that they can correctly identify relevant contexts in their particular research activities.

Simple statistical techniques are presented for filtration of meaning on unstructured text together with presentational paradigms from the domain of information visualisation which make language easier to compare, contrast and understand within the collaborative research environment. The novelty here rests both in extensions to existing presentation methods for corpus data and in the creation of an iterative search workflow for linguistic analysis of a living legal corpus in a workflow tool aimed at lawyers (and not linguists) for the first time.

The work in this thesis has been done with an overarching goal that someone without technical or data analysis expertise can understand what is being shown. In Chapter 7, an evaluation of the LARC software involving several user groups of lawyers and law students is presented and the conclusions from this exercise are then discussed. The discussion (in Chapter 8) presents important points for future

work and for refinement of the techniques and software that have been proposed.

1.2 Thesis statement

This thesis presents the hypothesis that *an integrated legal research platform which provides contextualised techniques for analysing, comparing and contrasting the language within a large corpus of legal information can help to engender effective research skills in law students and early-stage practising lawyers. This can be achieved by allowing for the guided refinement of information needs and the filtration and ranking of results that are returned from initially simple user-generated queries in real time.* It investigates the role that current computer-based research tools play in perpetuating poor legal research skills and how design decisions can be changed in order to address the issue. This work is founded on an analysis of how other domains deal with broad sources of written text using computers.

A pedagogical imperative of first training modern lawyers in the skills of critical thinking and linguistic analysis for themselves is identified. This requires a switch in the almost singular focus in legal education from acquisition of knowledge to a dual perspective where the application of that knowledge in real or quasi-real legal environments is equally valued, is compulsory and is enshrined in curriculum design. Techniques for text selection, filtration, display and manipulation from the domain of corpus linguistics are applied in an integrated manner to legal sources for the first time. This thesis thus attempts to illustrate how specialist computer-based tools can further educational priorities for newly-qualified lawyers and legal practitioners in order to enhance the skills and improve the confidence of trainees once they complete their legal education.

1.3 Scope

Corpus linguistics is a young science which developed from the mid-1960s as the availability of mainframe computers in universities became widespread. The potential of these computers to play a role in the analysis of language was quickly identified by first-generation corpus linguists like John Sinclair. Fundamental processing power became available at the same time as optical scanning facilities and early character recognition approaches also started to mature into practical systems. Sinclair and a small group of other linguists saw the potential of computers together with text scanning technology to change the state of the art in

linguistics. They developed new corpora based on both transcriptions of spoken language and written materials which very quickly deepened understanding of language use.

These resources could now be living and evolving accounts of how people write and speak to one another. Thanks to the efforts of the early corpus linguists, a rigorous science for the creation and interrogation of representative corpora using computers quickly became established. Their work incorporated how to select texts for inclusion in a corpus; whether to consider whole documents or fragments of text; how to clean and annotate texts for linguistic purposes; how to transfer these annotations into the digital domain; and how to present search results in a way that a linguist could understand and under ranking and selection criteria that were transparent and reliable.

It is true to say, however, that corpus linguistics was and remains a conservative science. The work that was done five decades ago to allow for interrogation of corpora using computers has not evolved at anything like the same pace as the field of information technology. Whilst fundamental developments like the Internet and the World Wide Web slowly gave rise to ideas like using *the web as a corpus*, this was hindered by lack of permission, copyright norms and problems about how complete and accurate online sources truly are.

Today, then, we have a science that still uses many of the standards set in the 1960s because those standards effectively answer the immediate needs of linguists and lexicographers for the limited purpose of analysing language use in an academic sense. Presentational norms like the *KeyWord In Context* display in corpus software, for example, feature lines of text that are eighty characters long because that was the standard width of a bale of continuous paper for computer printers. As recently as 2005, John Sinclair conducted the majority of his corpus queries using a text-based interface through Telnet sessions with a mainframe at the University of Birmingham [77].

If corpus linguistics is a conservative science then the legal profession and legal education are even more traditional and resistant to change. This author studied law at a major English university two decades ago and, despite the universality of computing resources at that time, legal research was still conducted by visiting a law library, speaking to expert librarians, consulting printed citation indexes classified by subject area and then retrieving and reading printed volumes like

the *All England Law Reports*. Computer-based access to these legal materials has developed since and has perhaps instigated the most profound changes in legal education.

Law firms and practising lawyers were the last to drop printed law libraries and their shelves of bound collections of law reports. This is happening quite quickly now and more legal organisations close their physical libraries every year whilst using tools like *LexisLibrary* [115] and *Thomson Westlaw* [182] online as a complete replacement. It is these computer-based tools, together with other products from smaller companies like *JustCite* [46] and *JustisOne* [90], which form the existing ecosystem which LARC seeks to disrupt.

The aim of this thesis is to start bringing together legal research and empirical linguistics in a manner which can benefit participants in both areas. There is a need to update the norms in corpus linguistics so that the very valid and useful developments in this field can move outside a linguistic perspective and can be effectively applied in other areas. Legal research and the ways in which lawyers work with computers to prepare cases needs to change in order to address ever-increasing problems of research skills deficits and information overload as greater and greater numbers of cases are reported every year. If there is a significant problem with the research skills of law students and early-stage legal trainees, it is reasonable to look to computer-based and online platforms for first steps in meeting the challenge because they are now the dominant method for information discovery and assimilation in the profession.

1.4 A note on methodology and the role of BAILII

The requirements analysis that underpins this research is triangulated from three separate studies of legal working practice. Firstly, a contextual inquiry was conducted with students who were preparing moot cases for presentation in a simulated court environment. An educational context was chosen because previous work shows that collaboration techniques, computer-supported tool choice and preferences are formed largely when students are training to become practising lawyers. It was then necessary to understand how the initial findings from the student group related to professional practice. To this end, a set of interviews was conducted with solicitors who were also involved in judging moot cases. This meant that the participants had knowledge of current legal practice in

both the profession and in education. Finally, it was desirable to understand how generalizable and accurate the findings in education and practice were across different areas of the legal profession. An online survey was designed and distributed which evaluated collaboration techniques and the role that computers play in that to a larger cohort of practising lawyers with different specialisms.

The University of St Andrews does not have a law department and, as such, the contextual inquiry was arranged with a senior academic at another Scottish university which is close by. The sessions were organised after an initial contact was made with the external university through an email to the Dean of the law department. A pool of eight participants for the enquiry was initially selected and this split naturally between existing mooted teams which were preparing cases for various mooted competitions at the time. The interviews with solicitors proved harder to arrange because initial emails to law firms in the vicinity of St Andrews went unanswered. As an alternative, the Law Officer at St Andrews was contacted. He forwarded details of the research project to a law firm in Central Scotland which he had worked with previously. A senior solicitor at this firm then helped to arrange a series of interviews with various colleagues over the course of a day. The broader survey of professional legal practice was also distributed through the Law Officer, who sent a link to the online questionnaire to various firms that he had existing relationships with.

Once the LARC software had been produced and an evaluation of its strengths and weaknesses was required, an email about the software which explained the framework for a remote evaluation session over the internet was sent to the dean of every law school in England and Wales. From this cohort, we received ten expressions of interest. Once the evaluation questionnaire was prepared and made publicly available, the link to the evaluation and to the prototype software was distributed amongst the ten interested law schools. From the initial pool, a total of four participants from the different schools completed the evaluation. In an attempt to boost the response size, the Law Officer at St Andrews sent the evaluation link to law firms in the area and a further five participants were recruited and provided feedback from this tranche. There was difficulty in recruiting participants at every stage of the research, especially from commercial law firms, and this is reflected in the small but adequate size of the ultimate response pools at each point in the thesis.

The work undertaken here would not have been possible without the assistance

of the British and Irish Legal Information Institute [111]. This organisation was contacted by email at the very start of the project, before there was a clear direction for the work. The Chief Executive agreed to grant a license for the use of their case law database in the LARC software. This license was only valid within the School of Computing at St Andrews, for research purposes, by the author of this thesis and his supervisors.

1.5 Contributions

This thesis contributes to the understanding of how English lawyers train to present cases in court and the role that computers play in the process of legal research. It examines problems and barriers to effective working practice for lawyers which are introduced by computers both during their education and during their time as early-stage trainees in law firms. A central argument is that new software developments aimed at lawyers need to be based upon open access legal data from the English legal system. Tactically, it is suggested that the development of open source platforms and tools to work with this open access data will encourage a sustainable diversification of legal information products by lowering barriers to entry and the total cost of ownership of any one solution.

An initial prototype open source platform for legal research is offered. This software incorporates algorithmic and presentational paradigms from the domain of corpus linguistics into the legal sphere in order to deliver an initial online environment which enables collaborative legal research based upon the detailed linguistic analysis of previous cases.

1.5.1 The move to open knowledge: open data and open source

The idea of *open knowledge* is based upon the publication and subsequent exposure of *open data* to users in ways which make that data useful to them. It relies upon content and information which is freely available to anyone who is interested in it. Openness encompasses the freedom to use, re-use, modify and re-distribute data without any legal, technological or social restriction. The “Open Definition” sets out the key principles which define openness in relation to data and content. This definition holds that “*knowledge is open if anyone is free to access, use, modify, and share it - subject, at most, to measures that preserve provenance and openness.*” [176]

In order to be open, data must be available as a whole and at no more than

reasonable reproduction cost. The content must be published in a convenient and modifiable form and this is usually facilitated by dissemination via the Internet. There must be provision for reuse and redistribution of the data which foresees and allows intermixing it with other content. Finally, there must be no discrimination in either the system of publication or the conditions of use. Commercial and non-commercial applications are equally valid. Everyone should be able to make use of the data for their own ends. There are various licensing schemes and agreements which are compatible to different extents with the definition of open data. These contain a range of terms for content attribution, subsequent sharing of derived or modified products based upon open data and a requirement to keep resultant products licensed openly. The most liberal open data licenses are the *Open Data Commons Public Domain Dedication and Licence* [143] and the *Open Data Commons Open Database License* [142].

The related idea of *open source* usually refers to software. It means that the source code for that software is openly available, thus allowing for inspection, modification, enhancement and forking, or the establishment of derivative software on an original codebase. The software may be redistributed freely. Most open source software is free of cost, but some applications do carry licensing fees. There are a wide variety of legal licensing frameworks which the authors of open source software can choose between in order to dictate how access to their code is facilitated and constrained. In general, open source licenses grant users permission to use the licensed software for any purpose they wish. Some open source licenses, which are known as “copyleft” licenses, stipulate that anyone who releases a modified program based on the source code of an original project must also release the source code for that underlying program alongside it. Moreover, some open source licenses stipulate that anyone who alters and shares a program with others must also share the original source code without charging a licensing fee for it. Common licenses for open source software include the General Public License (GPL), the Lesser General Public License (LGPL) and the Berkeley Software Distribution (BSD) license [56].

“Creativity flourished there because the Internet protected an innovation commons. The Internet’s very design built a neutral platform upon which the widest range of creators could experiment. The legal architecture surrounding it protected this free space so that culture and information - the ideas of our era - could flow freely and inspire an unprecedented breadth of expression. But

this structural design is changing - both legally and technically. This shift will destroy the opportunities for creativity and innovation that the Internet originally engendered. The cultural dinosaurs of our recent past are moving to quickly remake cyberspace so that they can better protect their interests..."

[112]

There is a consistent argument throughout this thesis that legal data under the English legal system should be made available on an "open data" basis. Some limited progress has been made here thanks to the publication of case transcripts by the courts themselves on the web. The British and Irish Legal Information Institute [111] was established in the early years of the current century in an effort to improve general access to legal information. This is a significant step which is of central utility to this research because the products that BAILII have published form the basis of the legal information used in the software which is presented here, under a bespoke licensing agreement for the research project. These sources remain relatively limited in scope and disparate at present, however. It is also the case that the standard licensing terms used by organisations such as BAILII are not sufficiently liberal to render their publications "open data". Modification of the sources and their inclusion, parsing and presentation in part or as a whole through derivative systems is prohibited to most users. In order for this situation to change, the movement towards open knowledge needs to gain impetus for both philosophical and practical reasons. This research would not have been possible without a bespoke licensing agreement with BAILII since the text mining, presentation and summarisation techniques implemented in the software require extensive modification of and extrapolation from the underlying sources.

The idea that a person should be able to easily establish the current state of the law and thereby to understand their rights and obligations under it demands that it be possible for them to simply determine what the law is. The centrality of precedent in the English system means that a fulsome appreciation of the law requires access to a broad range of information sources and to both historical and contemporary legal case reports and legislation. The currently dominant model of closed publication of legal information which can only be accessed by the general public on an onerous subscription basis or through a solicitor cannot be said to adequately fulfil the general requirement of access to and understanding of the law for everyone regardless of means. Thus there is a clear need to move towards publishing legal data on an open basis in order to improve access to the law and

to justice.

Secondly, this thesis presents an argument that diversification in the legal software ecosystem will be greatly aided by a focus on open source software development. This is essentially a tactical decision because it is very difficult to compete with large established software companies in this sector on a commercial, closed-source basis. The providers of legal software also tend to be information publishers in their own right and the dominant companies here have been offering products to law students, lawyers and academics for many decades. Therefore there is a severe disparity in resources and experience between these existing companies and any new companies which attempt to enter the marketplace. New entrants have a difficult task to secure access to legal information in the first place. It is also likely that the amount of development effort which is required to produce a competitive legal research product would require large scale financial investment even before any new software could be released. The open source model, by contrast, offers a possibility to distribute development amongst a large and diverse pool of developers and legal experts on the basis that they are interested in the work and in the release of a final product. Established open source developments like LibreOffice [58], Ubuntu Linux [181] and the GNU Image Manipulation Program (GIMP) [167] (amongst others) demonstrate that it is possible to create and to support large and complex software products on an open source basis. Thus the open source focus in this thesis is essentially a decision which is motivated by a desire to release products for lawyers which quickly and comprehensively compete with established commercial platforms with minimal direct financial investment.

1.5.2 Main research contributions

Figure 1.1 gives an overview of the scope of this thesis, the main theoretical areas considered and the placement of novel contributions within the overall design space. The fundamental contribution of this research is an open source, corpus-based collaborative environment for legal research. This is built from open source software components and upon open access legal data from the British and Irish Legal Information Institute. It prioritises contextual linguistic analysis as a method for imparting a detailed understanding of legal precedent to system users. The move towards providing and exploiting open access data and open source software in the legal domain is slowly gathering momentum, particularly

1. INTRODUCTION

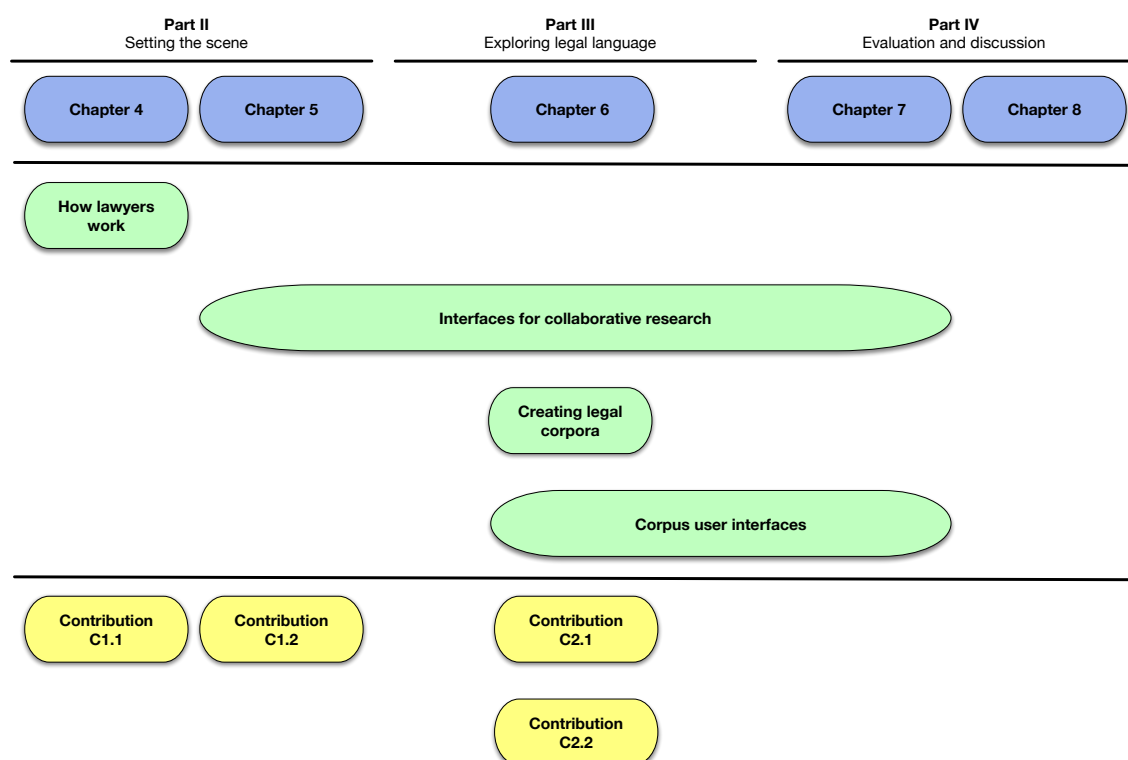


Figure 1.1: Overview of the structure of this thesis, the main topic areas concerned and key contributions.

in the United States [173] and the European Union (under the OpenLaws project [193]) - but also in England [81]. This thesis presents the first entirely open source development of a research software platform for lawyers.

“Individual parts of the puzzle are coming together, with stronger legal information institutes, more government commitment to publishing statute and case law databases in open formats, greater willingness by commercial publishers to experiment, and more law firms investing in open publishing of expert commentary. What is still needed is to bring these together with a social layer, to create a community which sees collaboration as more than simply checking LinkedIn every month. The value of open access is in creating a better educated legal community and general public.” [124]

The design goals and priorities for development here are initially identified through a novel contextual inquiry exercise which specifically looks at the benefits and problems of computer-based legal research tools for the first time. The work is based on the principle that legal language should be presented to expert users at

different stages of their development (legal students and trainee qualified lawyers) so that they themselves can learn to make objective judgements about relevance, result quality and suitability for use in particular novel situations. It also enshrines the idea that the jigsaw of generalist and specialist tools that is currently used for legal research presents a fragmented user experience and is counter-productive to effective working practice.

The new system also provides detailed progress reporting and work audit capabilities so that teachers and senior lawyers can evaluate and direct the activities of less experienced colleagues and students. This thesis includes an evaluation of the new research platform by groups of legal students and practitioners. Benefits, problems and future development of the approaches taken are finally discussed. The novel contributions of this research can be enumerated in more detail as follows.

C1.1 - Work analysis through contextual inquiry of how legal students at university and groups of qualified lawyers use computer-based research tools:

The contextual inquiry, interview and survey work provides a detailed description of how existing computer-based tools are used for legal research in universities and in practice for case preparation. The benefits of current online resources are considered whilst several important usability barriers are identified for the first time.

C1.2 - An open source collaborative environment for legal research: A new tool for legal research in groups which facilitates both synchronous and asynchronous collocated or remote collaboration is proposed, described and prototyped. This is a single-context, integrated software platform designed for legal research and is the first dedicated development of its type.

C2.1 - A representative living corpus of key English legal sources: To date, the only corpora which support linguistic analysis of English legal texts are archaic, static and time-limited. This contribution is a collection of key modern sources which are relevant and which describe the state of the law as it currently exists. The corpus is constructed carefully using accepted techniques from corpus linguistics for source creation and maintenance. A toolchain is provided which enables the new legal corpus to be a living resource which can be easily updated.

C2.2 - A user interface for legal corpus interrogation: An interface for legal corpus interrogation which is designed to facilitate knowledge synthesis in groups that are not composed of linguists or technical experts is proposed, described

and prototyped. This interface moves the state of the art in corpus evaluation tools forward by providing an easy-to-use but powerful system for the linguistic analysis of legal texts which does not rely on knowledge of query languages and which integrates seamlessly into a practical tool for legal research. Several novel applications of ranking and selection algorithms are proposed and implemented.

1.5.3 Additional outcomes

In addition to the main contributions, this thesis also includes smaller research outcomes which can be described as follows.

C3.1 - An automated citation layout for legal cases and legislation which integrates sentiment analysis: A hierarchical tree visualisation for the display of cases cited in a particular case, cases that cite a particular case and statutes that are cited in a particular case is proposed and implemented. This tree features on-demand node loading for citation exploration. It takes the state of the art in citation visualisation in manually-curated legal tools like *JustCite*, automates and simplifies the paradigm in order to make results more accessible, the layout more space efficient and the referencing of cases in different contexts easier to interrogate.

C3.2 - An alternative visualisation for KeyWord In Context result sets: This research proposes a new focus on linguistic variation so that different uses of language can be understood faster and with more clarity. It is suggested that the existing KWIC paradigm from corpus linguistics is ineffective at exposing variation outside the query node. A new hierarchical visualisation called ChoiceTree is proposed which is specifically intended to allow for the exploration of linguistic variation around an original query so that different word senses and phrases can be identified and disambiguated easily.

1.6 Thesis outline

This thesis is composed of four parts, each in turn comprising multiple chapters. The first part is 'Part I: Background and Related Work' which gives a general overview of the legal domain and key legal issues which are relevant to the rest of the thesis. It starts with 'Chapter 2: Background', which gives an overview of the domain of interest and describes fundamental parameters within which the rest of the research and the thesis as a whole operates. The next chapter,

'Chapter 3: Related Work', provides an overview of the application of computers to legal research, the state of the art both in this area and in the area of linguistic analysis using computers. It also describes visualisation paradigms that have been developed both for corpus interrogation and for elucidating the state of the law.

The second part of the thesis, 'Part II: Setting the Scene', describes research to understand how lawyers use computers in their work at the moment. It proposes an initial system for facilitating collaboration between groups of legal students and practising lawyers as they prepare cases. 'Chapter 4: Contextual Inquiry' recounts the work analysis with lawyers, trainees and students which forms the basis for the rest of the thesis. It elucidates a set of barriers to effective working practice which have been identified as problems with existing legal tools or areas where current platforms do not provide adequate solutions. 'Chapter 5: The LegAI Research and Collaboration platform' discusses in detail how the barriers previously identified can be addressed. It presents a set of development priorities which are enshrined in an initial open source prototype software application for use by lawyers.

The third part of the thesis, 'Part III: Exploring Legal Language', discusses the importance of linguistic analysis using computer-based corpora. It describes a new user interface for use by lawyers which is designed to foster a detailed understanding of language. The goal is to enable users to better identify the significance and relevance of specific and defined legal precedent from an ever-increasing quantity of textual information. 'Chapter 6: Integrating language search' concerns the development of a modern, living legal corpus by observing and extending the principles of source creation that have been laid down in the domain of corpus linguistics. It discusses implementation of an interrogation interface for this corpus for use by lawyers as part of their research workflow.

The last part, 'Part IV: Evaluation and Discussion of Results', presents an evaluation of the effectiveness of the LARC software based upon user group testing with legal students and practising lawyers. This part also discusses and summarises the findings of the previous parts and provides an overview of their significance in the context of this thesis. 'Chapter 8: Evaluation' discusses the design and deployment of the evaluation exercise and presents its results. 'Chapter 9: Discussion' summarises the results of the foregoing chapters and shows their relevance to the overall research question proposed in the thesis. The last chapter, 'Chapter 10: Conclusion', summarises the presented work, points out

its significance and shows potential future directions and follow up work that can be undertaken based on the outputs and contributions from this thesis.

1.6.1 Research question

The key research question that is posed in this thesis is:

How can methodologies and approaches from corpus linguistics be integrated into a research platform for lawyers?

In answering the key research question, a process will be followed in order to establish a grounded approach to applying corpus linguistics in research software for lawyers. This is done in order to ensure that the proposal for a new legal research platform is based upon the needs and requirements of law students and practising lawyers. It is also designed to answer any shortcomings that are identified in existing legal research software, where possible.

The process followed in the rest of this thesis which culminates in answering the key research question is enumerated in Table 1.2. The process and the answers that are arrived at in the course of following it will be revisited and the conclusions to each stage summarised in Chapter 8 (Discussion). The main research question will also be answered, in the context of all that has been learnt, in Chapter 8.

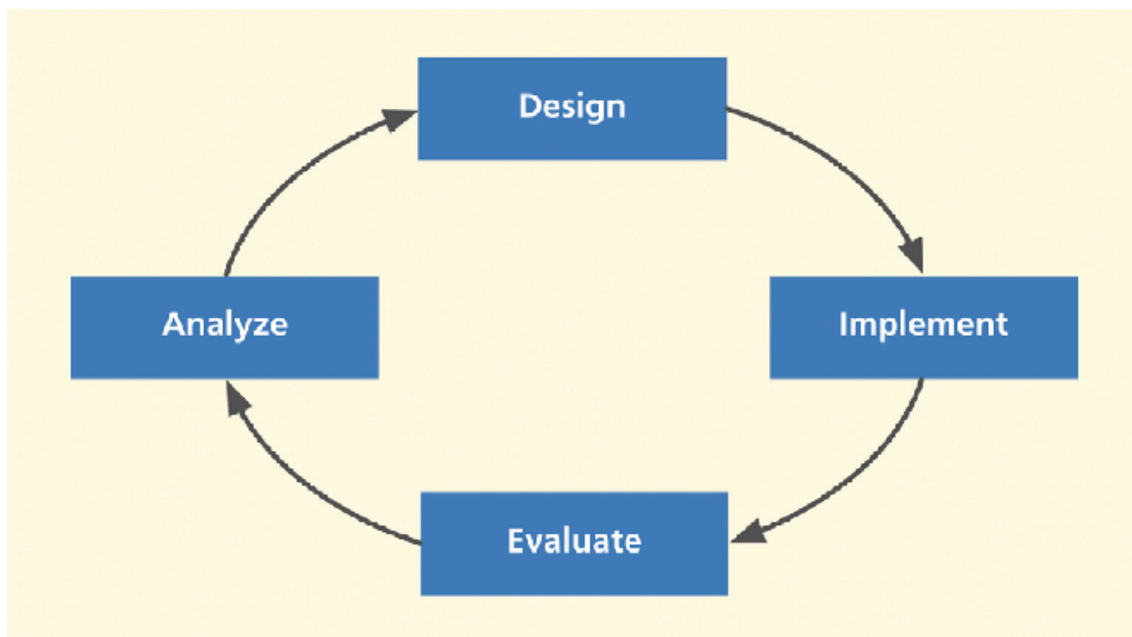


Figure 1.2: Overview of the UX lifecycle as presented by Hartson and Pyla in [71]

Step in process	Purpose	Lifecycle stage
P1	Consider how lawyers work in conducting legal research for case preparation	Analyse
P2	Consider the role that technology plays in facilitating or hindering effective working practice	Analyse
P3	Consider how open source software can be implemented as an integrated platform for legal research	Design
P4	Consider how synchronous and asynchronous collaboration on legal documents can be supported by technology in an auditable manner	Design
P5	Consider how search interfaces for linguistic content can be designed to target end-users who are not linguists or computer programmers	Implement
P6	Consider how usable the LARC software prototype is according to standardised user experience metrics	Evaluate
P7	Consider the feedback about LARC from lawyers and law students which should form the basis of future work on the platform	Evaluate

Table 1.2: The process that will be followed in order to answer the research question that is posed in this thesis.

1. INTRODUCTION

The process outlined above is designed to answer the requirements of the iterative user experience (UX) lifecycle, as proposed by Hartson and Pyla in [71]. The **Lifecycle stage** column in Table 1.2 ties the process followed in this thesis into the appropriate stages of this lifecycle.

PART I

BACKGROUND & RELATED WORK

BACKGROUND: LAW

2.1 Introduction

In this chapter, several important principles of the English legal system are introduced. This serves to contextualise the research and to identify key theories and customs which impact upon how legal research using computers is undertaken. The development of English common law is discussed and it is differentiated from other jurisdictions that are based on interrogatory civil law frameworks. The idea of “the rule of law” is considered in the context of a common law system of justice. The key sources of law are described, covering both legislation and case law.

The question of judicial interpretation of legislation is discussed with reference to both textualist and intentionalist theories. This is important because it dictates how methods for linguistic analysis of legal texts should be conceived and implemented in search and retrieval systems. The manner in which the doctrine of precedent imposes structure and hierarchy on case law by creating dependencies from new cases back to previous decisions is explained. The chapter also covers the special nature of legal language. It describes several ways in which the language of the law is different to ordinary language and considers the implications of these characteristics for linguistic analysis using computers. Finally, the chapter considers the traditional model for legal education and training and then describes modern efforts to shift emphasis away from doctrinal approaches to simulations.

2.2 The development of English common law

An understanding of how the English legal system works is key to **Contributions C1.2 and C2.1** of this thesis. The mechanics of justice in this jurisdiction make it essential that legal information be made available on an open access basis because it is fundamental that any citizen is able to investigate and to know the state of the law. The English legal system operates through a framework of “common law”. This has come to mean an adversarial system of justice. Both the plaintiff and the defendant or each of the opposing parties are represented and cases are decided through argument between the sides with an impartial judge overseeing the process. In fact, the historical scope of common law is narrower. It originally meant only that the whole of England should operate under one common system of laws. Although the United Kingdom today is a unitary state in international law, it was composed until 1536 of three different legal jurisdictions - England, Scotland and Wales. The Welsh began following English common law in the sixteenth century but the Scottish legal system remains somewhat distinct to this day [135].

This thesis is predominantly about the teaching, learning and practice of English law. The English model gives a broadly-applied basis for the legal systems and practices of other countries including the United States, Australia and Canada. Although these countries have independent legal systems and judicial structures, the genesis of their legal frameworks is in the English common law.

“The concept of binding, national law is a recent one, which has provoked fierce opposition [to the application of foreign precedent] in those countries (notably the United Kingdom, the United States, France and Germany) having most closely adhered to it.” [62].

Most scholars date the development of English common law to the aftermath of the Norman conquest in 1066. Before this date, England had been following Anglo-Saxon law. This held that local customs governed most matters and the Church played a leading role in government. Crimes, which were often based upon blood feud, were treated as wrongs for which compensation was made to the victim. With the accession of Alfred the Great in 871 AD, local customs were codified in a series of “*dooms*” (the Anglo-Saxon word for laws). However, this codification had little impact upon the resolution of disputes because law and

justice were still devolved to local areas. There was no centralised system of justice and no consistent system for legal restitution. The “*dooms*” were heavily weighted to issues about land and property ownership [85].

After the Norman conquest of England and the Norman victory at the Battle of Hastings in 1066, William the Conqueror did little to change existing Anglo-Saxon law because he viewed himself as legitimate successor to the throne. There were also significant practical difficulties about changing the existing systems of local law and custom. However, the Normans started to establish a centralised administrative framework in England based upon the principle of continental feudalism. Through this process, a unified body of land law developed and was eventually applied throughout England.

The advent of feudalism also had an important side effect because, in many cases, literate clergymen acted as the administrators of land and the effective judges in disputes about property ownership. Some of the clergy were versed in Roman law and the Canon law of the Catholic church [32]. Through the centralised administration of land, therefore, the principles of Canon law came to be applied in English church courts. This reliance on the tenets of Canon law changed the popular interpretation of crimes as personal matters. Canon law held that there was moral guilt in the perpetration of crimes and that, as such, crimes were a public matter.

The advent of King’s Courts and land administration contributed to the gradual development of a common set of laws in England. Courts under William largely attempted to act in a compatible manner with existing Anglo-Saxon laws and local customs to the extent that they could be reconciled together. It was not until the accession of Henry I in 1100 and, even more importantly, that of Henry II in 1154 that the modern concept of a common law becomes recognisable. Henry saw that the existing system of devolved local justice was unwieldy and, crucially, presented an inefficient system for the collection of revenues from land duties and other taxes. He set about the centralisation of judicial structures.

The first people to be formally appointed as judges by the King were experienced officers of the royal court who had advised Henry on the settlement of disputes. A subset of these officials was tasked with travelling the country on a regular basis to hear disputes and to administer and examine matters such as the payment and collection of taxes by and from the local aristocracy. In this way, the

2. BACKGROUND: LAW

process of administering justice and the machinery of that administration were centralised and started to operate under an already-established canon of rules and regulations that covered both issues of land law and broader questions of crime and punishment. Over time, a system of writs was developed for civil cases which identified specific actions that could be taken in particular sets of circumstances.

As the system of writs expanded with the consideration of more cases, they became inflexible and impractical to administer. There were complaints that the mechanical application of writs in some cases resulted in injustice for the plaintiffs. Dissatisfied parties who could not meet the requirements of a writ that found against them started to petition the King for relief. These applications involved deciding whether to grant relief to the applicant on the merits of their individual case. Through these decisions, a common corpus of valid grounds for relief was built up. The resulting body of laws became known as the law of equity.

In 1215, King John (the son of Henry II) met with dissatisfied members of the aristocracy who had risen up against him. The barons presented the King with a list of demands in a document known as the Articles of the Barons. King John granted the Charter of Liberties, subsequently known as Magna Carta, an act that ultimately led to peace. Magna Carta established, for the first time, many of the key principles of the doctrine of the rule of law. The foundations for a centralised system of justice based upon common laws and a right to personal liberty except in cases where that common law had been broken was realised. Cases were to be evaluated on their merits rather than by writ and presided over by an impartial and qualified judge.

“No free man shall be seized or imprisoned, or stripped of his rights or possessions, or outlawed or exiled, or deprived of his standing in any other way, nor will we proceed with force against him, or send others to do so, except by the lawful judgement of his equals or by the law of the land. To no one will we sell, to no one deny or delay right or justice.” [83]

2.3 The rule of law

“...stripped of all technicalities this means that government in all its actions is bound by rules fixed and announced beforehand - rules which make it possible to foresee with fair certainty how the authority will use its coercive powers in

given circumstances, and to plan one's individual affairs on the basis of this knowledge." [73]

The contributions made in this thesis are important because they seek to advance access to and understanding of legal sources. The rule of law can be made more transparent through the application of technology such that legal practitioners and laypeople are better able to practice their profession or to understand their rights and obligations under the law. The modern idea of the rule of law has become so accepted, particularly in democratic societies, that relatively little effort is expended on defining what is meant by it, how it evolves or why it is an important principle. Most of the public would likely accept the term without being able to properly define it. Lord Goldsmith pointed to the oath required of Lord Chancellors as an indicator of the importance of the rule of law in the modern age [64] but the paradoxical lack of clarity about the meaning and extent of the idea has been highlighted by Lord Bingham and others [16]. One of the most referenced explanations of the rule of law is that given by Friedrich Hayek, quoted above. This concise statement of the principle is a good starting point from which to understand the ideal of central and supreme laws, a clearly defined legal system and an independent judiciary in our society.

Contemporary definitions of the rule of law owe much to the thought and writings of Albert Dicey, the eminent constitutionalist and jurist who was working in the late nineteenth century [48]. He outlined three main principles which underpin the rule of law. Firstly, regular law must be absolutely supreme and predominant such that it excludes the existence of arbitrariness, of prerogative, or even of a wide discretionary authority on the part of the government. This means that *"...a man can be punished for a breach of law, but he can be punished for nothing else"*. Secondly, everyone must enjoy equality before the law, or the equal subjection of all classes to the ordinary law of the land administered by the ordinary courts. Thirdly, the laws of the constitution are not the source but the consequence of the rights of individuals, as defined and enforced by the courts.

Joseph Raz [153] describes at some length how the concept of the rule of law has been expanded to cover the virtues of a democratic political system, personal and social equality, an adherence to fundamental human rights and various kinds of respect for the individual and the dignity of men and women. These, he argues, are extensions to the fundamental principle which are unhelpful to the task of

arriving at a modern definition, notwithstanding the merits and validity of many of the broader ideas in themselves. He argues for a narrower and less political interpretation of the rule of law: firstly, that everyone should obey the law and should be ruled by it, and secondly, that the government should be ruled by the law and should also be subject to it. Implicit in these two ideas is the requirement that the law should be as clear as possible and as determinable as possible at any point in time [160].

The foregoing discussion demonstrates the vital importance of law and the legal system in our society. It also introduces the idea that law should be clear and accessible not just to lawyers but to members of the general public. If everyone is bound by the law, then they must be able without undue difficulty to find out what it is. In reality, of course, establishing the state of the law and its impact upon a set of circumstances involves taking advice from a solicitor or other legal professional who is trained in the subject and is able to navigate the textual universe of legal output with some skill. It is therefore in the interests of everyone that lawyers and legal advisors are educated effectively and have at their disposal the best possible skills and tools with which to examine issues and then to render legal advice that we can depend upon. The immediate sections of this thesis now examine how the law is created and structured to make this possible and how modern tools for legal research fit into the picture.

2.4 Sources of law: legislation and its judicial interpretation

“At the present day the most powerful instrument for legal change in the hands of the State is legislation.” [148]

From Chapter 1 of this thesis, **Contribution C2.1** seeks to bring together sources of statutory law and sources of case law in an integrated environment for legal research. The roles of both main sources of law in the English model are considered in this section. Legislation, or statutory law, is a body of law which has been made and enacted by a legislature, such as a parliament. A single piece of legislation is known as an Act of Parliament or a statute. Formulating new legislation and reviewing and amending existing legislation is the main business of the House of Commons and the House of Lords in the United Kingdom. Statute

is the central instrument that politicians have at their disposal to effect societal change, to authorise and to regulate private and collective activity, to sanction courses of action, to grant permission and to restrict liberties. Hayek argues that legislation originates from the necessity to establish general rules of conduct and to organise society. He also maintains that private law established through legal cases is rationalised and transformed into public law through the embodiment of principles already decided into appropriate statutory instruments [72].

Roscoe Pound argued for a fuller appreciation of the potential of legislation which is often refined at length by both politicians and committees of subject experts before becoming enshrined as the law of the land [150]. Particularly in the United States, there is an ongoing debate about the merits and deficiencies of the textualist approach to interpreting the principles enshrined in legislation. Some judges believe in a narrow prerogative to decide upon the ordinary meaning of words in statutes that are then applied in judgements. Others believe in a broader ambit under which it is the spirit and refined intent of legislation which must be ascertained.

“The most common way of distinguishing textualism from its principal judicial rival, “intentionalism,” purports to identify a basic disagreement about the proper goal of statutory interpretation: intentionalists try to identify and enforce the “subjective” intent of the enacting legislature, while textualists care only about the “objective” meaning of the statutory text.” [136]

The standard position in English law is that the lawgiver - Parliament - is the body that makes law, that only this body can change the law, but that it is the role of judges to interpret the words used in legislation in order to apply the law in the novel circumstances of individual cases. Therefore we have a dual system of law laid down in statutes passed by parliament and a concurrent body of precedent in the common law which provides binding authority on how legislation should be applied in practice. This is necessary because, just as with any other form of communication, legislative instruments use words which may have disputed or multiple meanings in different contexts.

The natural ambiguities of language give rise to contentious laws and competing interpretations of them. The role of judges in this context is to avoid a situation where badly-drafted or ambiguous statute leads to a failure of justice, such that a legislative instrument fails to render the results that were intended purely

because it has been written in a particular manner that is unclear in a given set of circumstances. There are competing ideas about how far judicial prerogative extends to interpreting the will of parliament in the United Kingdom as well. The literal approach holds that judges should look primarily to the words of the legislation in order to construe its meaning. Scope for examining the context in which legislation was drafted or the underlying purpose behind a statute are limited here. However, many lawyers and some judges advocate and practice a more purposive approach which gives significant scope for intentionalism in implementing statutory law [47].

In reality, there are various gradations of theory and practical approach which sit in between the literal and purposive approaches to legal interpretation. In the case of *River Wear Commissioners v Adamson* [1877] L.R. 2 App Cas 743, Lord Blackburn explicitly tried to define an extension to the approach of literal statutory interpretation. He said that “...we are to take the whole statute and construe it all together, giving the words their ordinary signification, unless when so applied they produce an inconsistency, or an absurdity or inconvenience so great as to convince the court that the intention could not have been to use them in their ordinary signification, which, though less proper, is one which the court thinks the words will bear”. At the other end of the spectrum, some judges have also attempted to define a much broader rule for judicial interpretation of legislation. This has been called the “mischief” rule. It concerns using previously-established common law rules in order to decide the operation of current legislation, thus indirectly placing a greater emphasis on existing law.

Most judges would probably select a method of statutory interpretation that was expedient and fair in the circumstances before them. In so doing, it has been established that reference can be made to intrinsic aids which facilitate the process of linguistic analysis and the interpretation of meaning. Intrinsic assistance may be derived from the statute itself, such that the judge can use the full content of a legislative instrument to understand a particular part of it. Thus the immediate context of a provision can be evaluated in order to guide its application on specific facts. The judge may also make reference to a limited set of extrinsic aids in their interpretative endeavours. They may use dictionaries to identify the meaning of non-legal words, textbooks to seek guidance on points of law and earlier statutes to determine the mischief that a later Act was designed to resolve.

There are also several presumptions that have been established which help to

elucidate how a judge can act in interpreting the language of statutes. There is a presumption against altering existing common law positions unless this is an explicit intention of a legislative instrument. Legislation in general should not have retrospective effect, thereby preventing the innocent actions of people in the past from becoming convictable through the passing of a new Act. The law courts also operate on the assumption that parliament does not intend to deprive a person of his liberty unless such a punishment is the explicit purpose of a particular law. The Human Rights Act 1998 also provides that judges must read all primary and secondary legislation from the United Kingdom parliament in a manner which is compatible with the European Convention on Human Rights. This means that, if a section of legislation has more than one possible meaning, judges must choose a version which is compatible with the ECHR.

The foregoing discussion of statutory interpretation under English law shows the importance of linguistic analysis as a skill which lawyers and judges can rely on in their work. Indeed, the final presumption which has been established for applying legislation states that words must take their meaning from the context in which they are used. This is called the Rule of Language. The rest of this thesis is founded on the importance of linguistic analysis and facility in legal training, legal practice and judicial reasoning.

2.5 Sources of law: the doctrine of precedent or *stare decisis*

“It is a basic principle of the administration of justice that like cases should be decided alike. This is enough to account for the fact that, in almost every jurisdiction, a judge tends to decide a case in the same way as that in which a similar case has been decided by another judge. The strength of this tendency varies greatly.” [39]

The English doctrine of precedent in its strict sense has grown stronger until modern times. The establishment of custom in itself relies to some extent on an adherence to precedents that have previously been set. However, Plucknett suggests that this process of reliance upon previous judgements initially came about quite unconsciously through an overarching desire to “*save trouble*” and to expedite the process of arriving at judgements in new cases [148]. Between 1250

and 1256, the cleric and jurist Henry of Bracton started to produce a systematised account of the state of the law which was based in part on the stipulations of statute and in part on judgements that had been passed down in previous cases. Because there was only a small body of legislation at that time, his work predominantly involved the collection of legal principle from case law [117].

It is clear that Bracton's use of cases interested his contemporaries. Other lawyers started to collect authorities for use in the same fashion. There was also a particular emphasis in Bracton's work on establishing the law on a good footing so that students and future lawyers could more easily and accurately elucidate precisely what the law was. It would take another two hundred years for court proceedings to be recorded in a regular and reliable format which included the judgement of the court in every case. By the mid-nineteenth century, a body called the Incorporated Council of Law Reporting was established with the express purpose of producing reports of significant legal cases in a standard format and this body continues to operate to the present day. The ICLR editorial and summarisation protocols give rise to additional materials like abstracts, keyword lists and headnotes in reports which make the process of quickly navigating and assimilating what can often be lengthy report documents much simpler. However, development of a broad knowledge of how previous cases had been decided only facilitated a precedent-based approach. The willingness of judges to be formally bound by previous decisions arose through custom and convention rather than by any more severe method of compulsion. The coercive system of *stare decisis* in England developed over time. This started with a vertical practice which meant that judges sitting in lower courts would follow the published decisions of appellate courts in the English legal hierarchy on questions of statutory interpretation and the application of the common law.

The main reason for the development of this vertical doctrine of *stare decisis* is probably judicial efficiency and conservation of effort. There is no need to consider a matter afresh when, through the interrogation of predictable and accurate reports of older cases, a clear line of existing judicial reasoning on the same topic can be established. The Judicature Acts of 1876 formally established that courts in England belonged to a proscribed hierarchy for the first time. The ultimate court in the land is the House of Lords (now the Supreme Court); the Court of Appeal is immediately inferior to it; followed by the High Court which is split into three divisions (the Queen's Bench division, the Family Division and the Chancery

Division); then followed by the County Courts and finally the lowest level courts, the Tribunals, which consider matters of employment law; immigration; freedom of information; land law; social security and child support; compensation for criminal injuries; mental health and pension appeals. Vertical *stare decisis* means, for instance, that the Court of Appeal is bound to follow precedents set down by the Supreme Court where previous judicial treatment of a principle on similar facts can be identified.

The vertical principle of precedent developed much earlier than the theoretically concomitant idea of precedent operating in a horizontal fashion. This means that courts can bind themselves in future situations based upon their previous findings in earlier cases.

“a decision of this House once given upon a point of law is conclusive upon this House afterwards, and that it is impossible to raise that question again as if it was res integra and could be re-argued, and so the House be asked to reverse its own decision. That is a principle which has been, I believe, without any real decision to the contrary, established now for some centuries.”

London Street Tramways Co Ltd v London County Council [1898] AC 375

The principle of horizontal *stare decisis* which was established in *London Street Tramways* persisted in a largely unchanged manner until the mid-1960s. At that time, the Law Lords decided that they would depart from the Tramways approach when and only when they felt that a failure to do so would create an injustice or would obstruct the development of the law. In reality, it has been suggested by legal scholars that previous precedent can and should be departed from in three situations: firstly, where a mistake has been made in the application of law; secondly, if fundamental principles that govern society have changed significantly since the original precedent was created; and thirdly, if personal values of principle and morality have changed in the intervening period. In combination, this sets a fairly high bar for overruling previous precedent [96].

“The delivery of administrative justice through tribunals comprises perhaps the largest part of the contemporary legal system. Tribunals annually determine a higher volume of cases than the combined output of both the civil and criminal justice systems. In 2010, tribunals heard 650,000 appeals,

whereas the equivalent volumes for criminal justice and civil justice were 223,000 and 63,000 respectively.” [54]

The lowest-level courts in the judicial hierarchy of English law - the Tribunals - have been growing in importance over a long period of time, thanks to the increasing number of cases which are presented and decided each year at this level. The role of tribunals was formalised, simplified and consolidated through the Tribunals, Courts and Enforcement Act 2007. This Act created a new system and context for tribunals which replaced a disjointed collection of separate tribunals with a single juridical structure. The reforming effort also heralded the completion of moves to embed tribunals firmly as a formal layer of the judicial system, deprecating their previous status as administrative bodies. Thus the 2007 Act effected a profound constitutional change.

Tribunals have not enjoyed a formal jurisdiction in the common law and they have therefore been seen to lack the power to set precedents. However, the second-level tribunals - which have now been subsumed into the Upper Tribunal under the 2007 Act - have created their own mechanisms for identifying important decisions which are to be followed by lower-level tribunals. This has been important because it creates guidance for the lower tribunals and provides judicial leadership within the system as a whole. Also, the higher courts have long attempted to respect the judgments of upper tribunals. In the context of asylum appeals, for example, the Court of Appeal has explicitly encouraged the immigration tribunal to give detailed guidance on the law in order to ensure a consistency of approach throughout compatible cases. Thus the tribunals have an important and increasing role in setting precedents that are respected both internally, within and between lower-level tribunals, and externally in the other tiers of the justice system.

The foregoing discussion serves to highlight both the evident hierarchy of court decisions from different levels of the English court system and the strongly coercive effect between individual judicial findings which is implied by the doctrines of horizontal and vertical *stare decisis*. For the purposes of this thesis, the important point here is that legal research needs to take account of the detailed and specific facts of individual cases which are similar to novel legal situations.

2.6 Research method: The focus on English law

“English law is important historically as a result of the British Empire, one of the two largest empires in recent history, alongside the French Empire. Part of the legacy is that its former colonies have modelled their legal systems closely on English law. Many countries which were formerly British colonies retain a system of common law (in which the development of and guidance and rules as to interpretation for the law are influenced by the input of the courts through precedent) and can look to the judgments of higher courts of England and Wales, particularly the Privy Council, for guidance on new or unusual issues. Similarly, the judgments of courts of other common law jurisdictions may also assist English courts in considering issues as they arise elsewhere.” [151]

A decision was taken at the outset of this research to focus on providing a prototype for a legal information platform under English law. As the quote above demonstrates, the English legal system has provided the basis for the legal systems of many other countries, thanks originally to the size and influence of the British empire. Former British colonies that have since become independent together with current commonwealth nations maintain legal hierarchies and systems of justice which are closely modelled on the English legal system. Although the legal system of the United States was originally derived in large part from English law, at both federal and state levels, it has diverged greatly from its English ancestor both in terms of substance and procedure.

Thanks to federal legislation in the US, initiatives like the Public Access to Court Electronic Records ecosystem have given a firm practical basis for open access to court reports and legislation. The vast majority of case law and legislation in the United States is now freely available online - a development which companies like Google have used to provide extensive legal search capabilities in their platforms which are available online to the general public at no cost. Given the importance of English law as the precursor to so many other legal systems, it is unfortunately still the case that open access under this system is much more limited and still seems to attract opposition in principle and practice from both government and from the existing purveyors of digital legal information. Although there has been some diversification of information providers in the English system, with new products like JustCite [46] being launched relatively recently, these new products still rely on commercial licensing agreements with established publishers

to furnish fundamental data.

“In [England], although legislation is openly accessible and reusable through the legislation.gov.uk platform, case law is problematic. Over the last century and a half, copyright in landmark and highly influential decisions of the courts of England and Wales has been claimed by various stakeholders representing rather different interests, from judges and court reporters to the commercial publishers who generate huge profits from annotating, enhancing and adding premium, feature-rich value to raw transcripts. More recent judgments are published and disseminated through numerous, separate court websites and in commercial databases accessible only to those individuals or organisations paying hefty subscription fees...Overall, case law cannot be said to be “open” in the same way that statute is open, consolidated and reusable.” [187]

Although using case law and legislation from the United States for this research would have been simpler, and the resulting platform would have been much more comprehensive in terms of the data available within it, it was felt that the development of a solution specifically for English law which operates on legal data from England could provide useful impetus as a proof-of-concept to advance the cause of enabling unfettered open access to case reports and legislation in this jurisdiction. Hoadley [81] states that the ultimate objective here “boils down to providing access that is free at the point of delivery to the text of every judgment given in every case by every court of record”. He supports this stance with reference to four tenets: the rule of law, the elimination of inequality before the law, promoting the effective resolution of disputes and creating legal transparency. These are all important considerations and they motivate the design and development of the LARC platform which is described in this thesis.

2.7 The diminishing role of domestic precedent

“When the six founding European states created the European Economic Community in 1957 they did so in the form of an international treaty (known as the Treaty of Rome) that was binding between them. That treaty also created the European Court of Justice. In an important ruling in 1964, the Court said that the states had agreed to limit their sovereign rights in the areas covered by the treaty and could not adopt national laws that were incompatible with

2.7. The diminishing role of domestic precedent

European law. This principle of “primacy” or supremacy of EU law has been accepted and applied by national courts including the UK courts. The Court of Justice does not, however, have any power to strike down national law; this is a task for the national courts. The national courts will, however, seek to resolve the conflict through interpretation. But UK courts are required not to enforce UK laws to the extent that they are incompatible with EU obligations.”
[179]

The system of domestic precedent under English law can be said to have diminished in importance recently thanks to the position of the UK as a member of the European Union. EU laws in areas for which the EU is responsible override any conflicting laws of member countries. The principle of supremacy, or primacy, describes the relationship between EU law and national law. It says that EU law should prevail if it conflicts with national law. This ensures that EU rules are applied uniformly throughout the Union. If national laws could contradict the EU treaties or laws passed by the EU institutions, there would not be a single set of rules in all member countries. The principle of the supremacy of European law over the decisions of English courts was first recognised in the European Communities Act 1972 and it has since been upheld and reinforced in the domestic courts. European Union regulations and directives also have “direct effect” in English courts. This means that parties to legal actions can rely on the provisions of EU law in court cases heard in domestic English courts. EU regulations take effect in the judicial systems of member countries as soon as they are passed whereas directives must be enforced domestically.

“A court or tribunal determining a question which has arisen in connection with a Convention right must take into account any...judgment, decision, declaration or advisory opinion of the European Court of Human Rights,...whenever made or given, so far as, in the opinion of the court or tribunal, it is relevant to the proceedings in which that question has arisen.”
(Section 2.1 of [180])

Furthermore, it is a general principle of English law that courts will seek, where possible, to interpret domestic legislation in compatibility with the international obligations into which the nation has entered. This is based upon the reasonable assumption that Parliament is not lightly to be taken to have legislated so as to

place the UK in breach of those obligations. Second, that general principle is given specific force under Section 2(1) of the Human Rights Act 1998, which is quoted above. This means that cases which involve questions of fundamental human rights must take into account the provisions of the European Convention of Human Rights, to which the United Kingdom is a founder and a signatory.

The Human Rights Act adopts a “dialogue model” which means that the domestic courts are invited by Parliament to indicate when legislation is incompatible with human rights and the provisions of the ECHR. Ultimately, Parliament decides if - and how - it will respond. The principle that courts must “take into account” any decisions made by the European Court of Human Rights when interpreting questions about fundamental rights applies only to the extent that the courts consider them to be relevant. This does not bind domestic courts but it does require them to accommodate and consider relevant European judgments - much as they do under the common law rules of statutory interpretation.

To add to this idea of the compatible application of domestic laws, Section 3 of the Human Rights Act requires anyone interpreting national law to do so in a way that is consistent with human rights - whether they are a court, tribunal or public authority. This applies to all legislation, including laws passed before the Human Rights Act came into force. Importantly, Section 3 includes the caveat that laws must be interpreted in a way that is compatible with human rights only to the extent that it is possible to do so. Section 3(2) of the Human Rights Act holds that this interpretative power does not affect the validity, operation or enforcement of any Act of Parliament. If an Act of Parliament requires secondary legislation to be made that does not comply with human rights, that secondary legislation will not be affected by Section 3. However, if the secondary legislation could have been drafted differently and could have complied with human rights, the courts can it strike down. This is because secondary legislation does not have the same status as primary legislation, which has been fully considered by Parliament.

The foregoing discussion highlights the fact that the importance and role of domestic precedent in English law has been reduced in modern times thanks to the adoption and creation of overlaying precedential authorities and courts from the European domain. One area of domestic law which has been particularly affected and changed in the European context is regulation for consumer protection. Consumer law changes from Europe have been described as “so all-pervading” [179] that it now forms the hardest area of law to disentangle, to maintain and

to promote domestically when the UK leaves the European Union. To know the state of the law today, it is necessary not only to understand the hierarchy and standing of national decisions and legislation in a particular area but also to have close regard to rulings from the European Court of Justice and the European Court of Human Rights. Indeed, the last few decades has seen a marked and progressively increasing tendency on the part of successive governments to dictate the direction and specifics of the law by enacting more and more primary and secondary legislation. The statistics reveal that there has been a long term trend for each government to legislate more aggressively than its predecessors. This has been most marked in areas such as criminal law, where forty Criminal Justice Acts have been introduced since 1997, and employment law.

In 2005 alone, some two thousand one hundred European regulations were enacted, all of which automatically take direct effect in English law without immediate domestic judicial intervention [172]. There has also been a move towards strengthening the role of the state as regulator in the domestic legislative agenda, notably in consumer law. Tony Blair's government, from 1997, was responsible for enacting 54% more laws per year than had been placed on the books by the Thatcher and Major governments beforehand. This increased tendency to "fix" perceived problems by legislating to change the law directly continues to be a feature of modern governance in the United Kingdom.

"We have lived in recent years in a blame/compensation culture which demands that somebody does something about every accident or bit of misconduct, and politicians and their departments feel obliged to react. Whether this is an issue of health and safety, consumer protection, discrimination, putting a regulation on the books or increasing a penalty to make a political point, even if not always followed up by adequate funding or enforcement." [172]

2.8 Language and the law

"Law is language...Laws are coded in language and the processes of the law are mediated through language...The language of the law is therefore of genuine importance." [60]

A central tenet of this thesis - which impacts on both the design and implementa-

tion of **Contributions C2.1 and C2.2** as enumerated in Chapter 1 - is that linguistic analysis of legal case law and legislation can be an important tool for fostering and promoting the skills of effective legal research. It is held that insufficient attention has been paid to producing computer-based systems for the interrogation of legal language. Existing platforms and software for legal research expose information through many pre-determined search facets but they fail adequately to expose the unstructured information in legal documents so that practitioners and students can make sound decisions about what is important content for their research tasks and which elements from the copious breadth of online information sources can be safely disregarded.

The technicality of the legal domain is manifested in various ways. Legal language frequently involves terms that are not part of everyday language. These terms have come to have a technical and contextual meaning within the profession which is often quite separate from any normal meaning ascribed to the words as a whole or in isolation that may be discerned from speech or writing in other domains. Some examples of specialised words and collocations which do not have a set meaning outside the law are offered by Gibbons [61], and these include “*codicil*”, “*deforcement*” and “*decree nisi*”. Legal writing also tends to overload normal words with specific and non-standard meanings. A good example here is “*contempt*”, which legally describes an act of deliberate disobedience or disregard for the laws, regulations and decorum of a public authority such as a court or a legislative body. Thus we have the associated ideas of contempt of court and contempt of parliament. Both of these concepts have at once a more specific, more technical and more onerous meaning than the standard definition of the word contempt itself.

The example given by Gibbons of the legal term “*decree nisi*”, which refers to a statement given by a court about the date upon which a marriage will end under divorce proceedings, has no common meaning outside of the legal sphere but it illustrates the importance and influence of Latin foundations in the common law. Another example here is the concept of “*res ipsa loquitur*”, which means literally that “*the thing speaks for itself*”. This phrase actually refers to a developed body of legal principle in the common law of torts. Negligence may be inferred under this doctrine from the very nature of an accident or injury in the absence of direct information about how the defendant behaved in relation to the causation of the injury.

Legal language reflects, therefore, a distinct microculture with technical terms and extended meanings for common phrases which are required in order to adequately describe legal concepts [132]. In a grammatical sense, the language of the law often features long noun phrases which are the vehicles for lexical technicality and complexity. A noun phrase includes a noun - a person, place or thing - and some modifiers which distinguish it. Researchers have found that legal language tends to feature a high density of nouns and noun phrases. It has been suggested that these constructions allow for the concise statement of complex ideas in a condensed written form [196]. The use of complicated noun phrases often leads to multiple possibilities for meaning and interpretation, and even to linguistic ambiguity or vagueness.

Pala et al [144] present a study of the language in fifty thousand Czech legal texts which again highlights the noun phrase as a key bearer of meaning. They identify almost four thousand distinct noun phrases from their corpus which carry a legal meaning that is different from ordinary language and can be said to be diagnostic of and specific to legal texts. Many of the noun phrases that are identified in that paper are repeated very frequently throughout the corpus. This indicates that there are lexical contexts in legal texts which occur often and which refer to well established norms, defined principles and standardised doctrines within the profession and the broader domain of the jurist. This meaning may not be of significance to someone who is not schooled in the law. As such, Pala's findings are an empirical demonstration of a thesis first promulgated by Brenda Danet a quarter of a century ago [45]. This idea has been extended to propose that complex syntactic structures are also used generally in legal documents, in combination with complex and compressed phraseology, in order to establish both the nature of laws and the conditions under which they apply.

"...increased complexity at the phrase level is usually accompanied by reduced syntactic complexity in the sentence or clause complex. Unfortunately, unlike scientific English, the language of the law appears to have the worst of both worlds, combining complex phrases with complex sentence syntax...The reason for this complexity appears to be that legal language is often trying to cover all possible combinations of conditions and contingencies." [60]

Mertz [133] argues that a key reason for the development of specialised and specific norms in legal language which deviate in syntax and style from general

linguistic usage is to allow for the creation, maintenance and codification of power relations in society. Legal language and its mastery confers an ability to challenge existing bases of power. The state uses language to impose its interpretations and its appropriations of physical and symbolic power and thus members of society must meet a certain standard of technical and linguistic facility in order to challenge and shift existing power relations. The language of the law represents the pattern and codification of power and jurisdiction in itself. That thesis is about the establishment of exclusivity, a high level of literacy and general educational attainment before one can be allowed to enter the legal sphere in order to engage with and potentially threaten the legal and theoretical basis of our society.

Maynard [131] takes this further by presenting research about the believability of people who give testimony in court. A key finding from his work is that jury members tend to find people with a high degree of technical linguistic ability and assurance to be more credible than those who speak in a simpler, everyday pattern of dialogue. Stygall [170] demonstrates the problems that laypeople and people with low levels of language skill would have in understanding their legal position, not just because terms used in statutes are legally overloaded, but because the ordinary meanings of many of the arcane words in themselves are convoluted or require advanced levels of comprehension:

“The difficulties with laypeople reading legal language are well documented and include problems ranging from the use of archaisms to terms of art, generally unknown to the lay reader. An additional problem is the use of vocabulary which has both a legal definition and an ordinary one.” [170]

An important claim in this thesis is that much of the utility of the legal profession, certainly for people in their initial interactions with qualified lawyers, is the sense of confidence that relatively disadvantaged parties can have and can feel through the very act of consulting a professional who has a proven record of legal understanding, educational attainment and linguistic facility. This, it is argued, could be compromised through an ill-considered rush to implement artificial intelligence and pre-determined ideas about how lawyers and their clients should interact. A central issue here is that many expert systems based upon machine learning and other approaches to the artificial intelligence problem are black boxes. It is not possible for people without a significant technical background to

clearly appreciate how software is making decisions about which legal information content is relevant or important in any given situation. As Radboud Winkels states:

“The user also needs insight into how the system works. Another term for acquiring insight into the working of a machine is the (re)construction of a conceptual model of the machine by the user. This model should correspond to the conceptual model that the developers had in mind when they designed the system. It can in many respects be [either] a black box or a glass box. To some extent a [system] is always a black box because it hides - and should hide - detailed processing. But hiding too much may inhibit acquiring insight into the workings of the system and result in [the development of] the wrong mental model.” [192]

The power relation between law and layperson and between lawyer and client should be an important source of guidance for the judicious application of technology in a manner which improves and facilitates the lawyer-client relationship rather than diluting or compromising it. One of the roles of legal actors is to present and translate an assured and highly literate legal and logical position in a form that clients, claimants, defendants and jury members can understand. Each side takes the arcane and legalistic language of the judicial process and convert it into a narrative of largely ordinary language for the purposes of comprehension and elucidation [76].

The difficulties of promoting comprehension between lawyers and laypeople has led to a movement in the last several decades to attempt to rewrite rules, regulations and statutes in plain English. The principles of the rule of law, equality of access to legal remedies and ease of access to the current state of the law imply that effort to demystify the profession by changing linguistic norms is important. Danet [44] charts the modern desire to achieve a closer correlation between legal language and ordinary language. She maintains that efforts by state and federal authorities to hold conferences on language reform helped to produce new versions of legal and bureaucratic documents. Parallel calls for the reform of legal language were heard in Europe at the same time. In England, the Campaign for Plain English asserts that “...it is possible to use plain English in legal documents. It does not mean sacrificing accuracy for clarity. The excuse that legal writing has to be complex to avoid misinterpretations does not stand up.” [177]

This is a problematic position, however. For one thing, although it is assumed that all “*legalese*” is difficult to interpret and to understand, there is little data, aside from anecdotes, to support the assumption or to elucidate the exact nature of the problem. As Charrow states [29], plain English campaigns lack the necessary empirical evidence of the extent to which legal language is not understood, nor is there enough data regarding those segments of the population - aside from lawyers and judges - that may not have problems comprehending legalese. The advent of computer-based tools for large-scale corpus analysis in recent years may change the way in which ordinary meanings can be determined and applied in plain statements of the law, aiding comprehension of complex language usage and scope both by the drafters of bureaucratic documents and by lawyers, but these initiatives are in their infancy at present outside the field of applied linguistics.

Another issue under the English common law system is that the very principle of precedent itself, and the idea that legal norms evolve through linked cases, means that the interrogator of a legal corpus will always eventually (and much sooner rather than later) come upon documents or case reports that do not attempt to and do not aspire to be easily understood by the untrained eye. A plain English legal system would take hundreds of years to become practically established as a result, even if it were limited only to redrafting statute, and the nature of ordinary usage will probably have changed beyond recognition in that time, making the utility of such initiatives questionable.

Finally, it is worth saying that, for all that legal language can be and often is very precise, and much of the judicial process concerns almost obsessively defining and extending rules through the language of new cases and the interpretation of legislation to fit new situations, that same language can also be deliberately vague. Vagueness and inexactitude are inescapable attributes of language and it is this lack of determinism which gives rise to an adversarial system of justice in the first place - there are two sides to every argument about what people say and what they mean by it. This exploitation of vagueness in language historically reaches its zenith in situations where a legislature and the judiciary are antagonistic towards each other. There is the maxim that “*statutes in derogation of the common law must be strictly construed*” [30], which is an idea that is fiercely applied in circumstances where the courts do not approve of a particular legislative instrument. More commonly, however, legislation is created to be deliberately vague in order to navigate a position of political opposition or discord. In recent times, commentators have

criticised the imprecision of the European Union (Withdrawal) Act 2018, initiating the United Kingdom's withdrawal from the European Union ([68], [129]).

2.9 The nature of legal training

"The history of legal education policy can be characterised as one in which periods of benign neglect have been interspersed with and punctuated by shorter periods of more or less intense navel-gazing. These latter interventions have not, on the whole, been initiated or led by those actively engaged in legal education and have seldom been actuated by a simple desire to build a superior system of education and training. Rather, they have been political engagements triggered by state or profession in response to perceived "problems"." [188]

Contribution C1.1 of this thesis seeks to establish how lawyers and legal students work as they conduct research. Many elements of working practice, including tool preferences and technological choices, are formed during training at university and law school. As such, an understanding of how legal education is designed and delivered is useful here. The modern educational framework for students of the law in England and Wales dates back only as far as the late nineteenth century. Before this time, university provision for professional education - in medicine, law and divinity, for example - has been described as "*virtually moribund*" [7]. The very idea of what a university is and what sort of education it should provide was a subject of intense debate in the mid-nineteenth century [137]. Until 1836, there was no requirement at all for foundational examinations or evidence of academic attainment before a student could be admitted to the Bar, for example. Webb [188] states that neither the professional legal bodies, like the Inns of Court, nor the major universities themselves seemed willing to change or to direct the development of legal training of their own accord. There was, however, a widespread concern amongst practising lawyers that the standards of newly-qualified lawyers needed to be more uniform and guaranteed through the implementation of some standard model for legal training.

"It has long been matter for general and just censure, that the Inns of Court - in theory the Law Universities of England - do nothing to promote legal education or the science of law. By eating a certain number of dinners in the hall of an Inn of Court, during a certain number of years, a vim man acquires

2. BACKGROUND: LAW

a right to be admitted to practise at the bar...[There has been] an acknowledged deficiency which has long been felt to exist in the education of English lawyers, in consequence of the entire neglect of the study of Jurisprudence..." [178]

The government finally acted on this discontent by establishing a formal commission of inquiry into the legal profession and legal education. This 1846 review became the first of six major reviews concerning the way that lawyers were educated and trained - the others reported in 1934, 1971, 1979, 1988 and 1997. The reports of these enquiries established how lawyers were trained at the time and to some extent tried to shape the nature of legal education so that it better met the changing requirements of the profession and the changing face of universities themselves. Collectively, they helped to consolidate English law as an academic discipline in its own right. They developed the idea of different stages within legal education, from a general undergraduate degree to professional qualifications which were suitable for either solicitors or barristers, so that different stakeholders could be assured that their requirements were met.

The route to qualification for the majority of students of English law has thus been largely standardised. There is a general academic stage at the outset which involves studying for a law degree or, potentially, a non-law degree followed by a conversion course at a later date. There is then the vocational course for qualification either as a solicitor or a barrister, followed finally by a training contract or pupillage. All undergraduate law degrees which may lead to qualification in the profession must qualify by covering certain foundation subjects that has to make up a defined portion of a three-year degree programme. The Universities and their law schools have a broad degree of freedom in deciding how to present individual subjects within a degree. Therefore degrees from different institutions may be quite different in character and in their depth of coverage of different subjects.

"At one level, there is a huge amount of freedom. One QLD might focus, in trusts and property, for example, on controversial policy issues such as homelessness or the charitable status of independent schools, while another could be concerned mostly with easements, conveyancing and the rule against perpetuities...At another level this is very restrictive. It is not self-evident that knowledge of trusts and land law is more important than non-foundation

subjects...And giving complete freedom to only one third of the curriculum limits programme design” [159]

In recent years, some commentators have published critiques of the modern law degree and qualification pathways. These focus on the idea that the foundation subjects and the broader nature of learning and teaching within universities and law schools has become too conservative and ingrained in their adherence to a model which derives very largely from the society of the early twentieth century. They suggest that the narrow range of skills which are required to be taught has led to a culture of doctrinalism, focusing on communicating knowledge about legal rules and structures and providing almost no broader framework of development - in social and political science, for example - for the aspiring lawyer [37].

Keyes and Johnstone [93] go further by identifying five tenets of legal education which persists from a model first standardised in the twentieth century. They argue that legal education follows a teacher-led model in which the priority is the communication of knowledge to students by subject experts in narrow areas of the law. This leads to a limited focus on the learning experience for individual students and a lack of priority is afforded to the skills of applying knowledge in a practical sense.

“Traditional legal education is almost entirely concerned with the transmission of content knowledge and, more particularly, with teaching legal rules, especially those drawn from case law...[T]here is no appreciation of the students’ intellectual development as they progress through their degree. The traditional law curriculum gives little express consideration to generic skills (such as oral communication, self-reflection, teamwork, computer skills and so on)...Students are given no opportunity in their formal education to learn from and with each other.” [93]

The latest review of legal education, the Legal Education and Training Review, reported in 2013. *“The most significant thing in the...report, which took over two years to produce following various delays, is its public acknowledgement of “considerable dissatisfaction” from students paying in pursuit of a career “they’re never likely to achieve”. Lest we forget, university tuition fees trebled last year. Meanwhile, the already high cost of legal training continues to rise. What that means, in the practical terms outlined in the report, is not the introduction of an aptitude test to determine who is allowed to go to law*

school - an option which the LETR team rejected for “diversity” reasons. Rather, a focus on opening up alternative, cheaper ways to become a lawyer is being encouraged.” [1].

In more detail, the Review seeks to change legal education so that it supports modern objectives in the delivery of legal services: to ensure an independent, strong, diverse and effective legal profession; to protect and promote the public interest; to support the constitutional principle of the rule of law; and to improve access to justice. Many of the resulting changes that are recommended in the report centre around diversifying the body of legal students and ensuring that a broader element of professional skills are taught during the course of qualifying law degrees. Most of these objectives have little to do with changing the nature of legal education. Although the report acknowledges some problems with the doctrinal approach to learning and teaching, there are few concrete proposals or recommendations about how to change legal education in order to address these. In response to the LETR, the Chief Executive of The Law Society said:

“Educational establishments which are privileged to deliver qualifying law degrees are leaving quality assurance to the profession. The feedback we are getting from law firms shows that graduates are lacking the skills expected of them when they commence employment.” [38]

For the purposes of this thesis, the finding that there is little emphasis in legal education on social skills, on teamwork and on using computers is of particular significance. In a context where legal research now requires advanced information technology skills for the use of online information retrieval platforms, the low priority of computer literacy and education in qualifying law degrees is of concern. The next section will consider elements of a movement proposed by some legal scholars and institutions to depart from knowledge acquisition as the heart of a law degree. They advocate a compulsory and central focus on activities such as mooting and other forms of the simulated application of knowledge.

2.10 Legal training: Simulation and mooting

“By simulation we mean any activity, be it assessed or non-assessed, which requires students to engage in tasks and challenges that replicate real life...[T]he use of simulation can enable students to gain insights into what will

be required of them in their professional lives through facilitating a “messy” problem-based approach to learning...Rather than just identifying the “correct” answer, simulation can involve the student in evaluating and reflecting on different ways in which the law...can be used to achieve the optimum result for a ‘client’...The academic law degree, by and large, has a curriculum that sets up boundaries between subject areas.” [169]

Contribution C1.1 of this thesis was partly designed to answer the requirements of law students at university who are engaged in mootings activities. Thus, it is helpful to understand how mootings and other forms of simulation fit into the educational environment for lawyers. Over the last four decades, there has been increasing acceptance within the legal education sector that traditional doctrinalist approaches to learning and teaching need to be modified and augmented with new techniques. Traditional methods which treat the student as a sponge that acquires knowledge from specialists in narrowly-defined areas of the law can be modified to include the development of a broader range of skills. Effecting this change is challenging because, as [169] points out, students themselves see the primary goal of a university or law school education as involving the communication of knowledge from subject experts and the subsequent replication of that knowledge in examination answers. There is also the challenge of suitably structuring and equipping law schools so that they are able to integrate simulation effectively into the broader curriculum [121].

Changing the nature of the law degree therefore sits within a much broader debate about altering higher education and professional training norms, and the expectations of students, across many different subjects, of which law is just one component. The primary motivating factor here is a financial one. Students pay ever-increasing fees to study for professional disciplines like law and they need some assurance that the skills which they develop during their studies are appropriate to the workplace. The legal profession itself also has financial and time expectations. They require that newly-qualified members of staff have well-developed skills of teamwork, information technology literacy, and legal research in order to reduce the burden on firms to plug gaps in these skills.

Simulation as a tool in legal learning can be interpreted broadly, as Maharg and Nicol chose to do in [121]. “[Simulation]...is construed as any heuristic that involved the simulation of any aspect of legal theory or practice within a legal education context

and for an educational purpose". There has been some important activity to assess the approach and to use it practically for teaching students over a reasonably long period of time. Much of the modern research in this area involves applying technology to create, facilitate and manage the environment of a simulation. Maharg and Nicol [121] identify nearly one hundred relevant academic papers from their systematic review of the domain where there is an intersection between legal education, simulated learning and technology. The chronology of these contributions shows that activity and interest along these lines has increased rapidly in recent years. The range of simulation platforms and technologies varies greatly. Since 1970, they have included search and retrieval activities based upon strategies backed by artificial intelligence, multimedia resources featuring audio and video presentations which encourage self-directed learning online using prompts, entirely virtual online environments for roleplay and transactional learning like Second Life, and more standard platforms for videoconferencing between "*clients*" and the "*legal professional*". The results of these studies show, according to Maharg and Nicol, that approaches for simulation of legal scenarios continue to sit uneasily within the custom and practices of traditional tertiary and professional legal education and training.

"[Teachers] require an infrastructure for a new employment category, including the recognition of educational and technical expertise and reward and career structures for this new category of personnel. There is a lack of coherence in method and particularly in evaluative methods." [121]

Part of the problem here is that technology-driven simulation and teaching in the law remains a "*shadow pedagogy*" for a number of reasons. The law incorporates and requires background knowledge from many different academic disciplines in order to be practised successfully and with skill. Reliable methods for teaching these diverse subjects using technology, wherever that technology falls on the spectrum from learning management systems to entirely virtual environments (which all necessarily divorce participants from anything like a one-to-one relationship between tutor and student), are difficult to deliver systematically. The traditional model of teacher-driven lectures and small-scale tutorials for real time one-to-few exploration of topics and legal problems has worked quite well for a very long time.

"[F]or learning to take place, the core structure of the conversational framework must remain intact in some form: the dialogue must take place somewhere, the actions must happen somewhere, even if it is all done inside the student's head." [120]

"If we expect students to collaborate in learning on a professional basis, what is the constructivist basis of that engagement? What are the drivers and blockers to successful learning in this network of relationships?...We want students to be involved in activities within legal actions, rather than standing back from the actions and merely learning about them." [122].

It is clear that teamwork and the ability to work in collaboration with others are key skills that law students need as they transition from education to training and practice. This finding is also reinforced by the above quote. Abdul Paliwala succinctly states the requirement for pedagogy which equips lawyers with the abilities to direct their own critical thinking and learning activities in collaborative environments. He argues that the expectations placed upon modern lawyers necessitate two fundamental tenets in legal education. *"Students should be given the tools and support to construct their own knowledge either individually or in groups; and student learning should reflect the community in which she is going to live and work - thus if lawyers work interactively in groups and negotiate, so should students."* [145].

Many of the modern approaches to simulation of legal scenarios involve creating more or less immersive environments for working in teams where those environments are constructed and facilitated using technology. However, simulation of the legal environment has a much more established history in tertiary and professional education for lawyers. Mooting is the process of addressing legal problems in the form of imaginary cases that concern real legal issues. Moots are argued by two student *"counsel"* on each side in front of a bench of *"judges"* in a simulated environment which is usually, under English law, modelled on the Court of Appeal or the Supreme Court.

The process is adversarial between the two sides, one of which often represents a plaintiff whilst the other represents a defendant. The judges of a moot do not have to be real members of the judiciary but it is often the case that retired members of the bench will take on this role. The judge can also be a lecturer or someone with a distinguished legal background. The two students on each side take on the roles

of senior and junior counsel, which is just what would be required in a real court. The two members of each legal team must work together in order to analyse a legal problem that is provided in advance and to build their competing cases for presentation to the court. Mooting is therefore a detailed form of simulation which encourages learning not just about the form and structure of the law but also the application of knowledge to a “real” scenario. It helps to foster other skills like collaboration, conforming to legal and court etiquette and public speaking. Smith notes that mooting may be a formal part of the curriculum in some universities and colleges but it is more common for the arrangement and execution of moot cases to be the responsibility of the students’ law society [191].

“All moot court judges may and should give counsel a hot time by interjecting questions and objections to the argument presented...The objection need not represent the judge’s real opinion; this is done in order to see how the student counsel responds.” [191]

As Lynch points out, some faculties take a much more sophisticated approach to the simulation by requiring written briefs to be submitted by each side and by employing other students in the roles of instructing solicitors, for example [118]. In these more developed forms, mooting may actually comprise a formal part of student assessment. Moots are often held as competitions with coveted and established prizes for the winning teams. Although there is clearly an established context for mooting in most educational systems, it remains relatively unusual for mooting to form a compulsory part of legal curricula or for the results to count towards the assessment and grading of individual students.

Mooting should also be differentiated from mock trials, as noted by both [109] and [191], since moots rarely involve witnesses and juries whereas trials attempt to recreate the complete court experience. Participation in the activities of the mooting society may be seen by students as a valuable addition to their résumés. However, it is often perfectly possible for a student to progress through their degree and to graduate without coming into contact with the mooting society.

“[I]t should be realised that it is not possible to criticise a law curriculum solely on the number of opportunities for mooting it gives students. Mooting involves a vast amount of administrative co-ordination as well as actual time spent assessing each individual student’s oral (and, in some instances,

written) submission and performance. In fact, mooting is perhaps the most time consuming activity which a legal academic can feature in a course and so the practical possibility of holding moots is limited more than is the case with most other forms of assessment, and is largely determined by staffing resources and student numbers” [118].

Many scholars see mooting as an effective method of simulation which encourages the application of knowledge. Lynch argues that it is “*impossible to moot successfully without interpreting and abstracting meaning from the vast amounts of case and statute law (let alone academic writings) relevant to the moot problem. Moots involve memorisation and retention of knowledge, but first the mooters must construct that knowledge from the materials that they will discover through their research*” [118].

It is the intersection of this knowledge acquisition through research and the subsequent synthesis and shaping of that knowledge which forms the crux of this thesis. The justification for applying technology to this process, the manner in which that technology is applied and for what purpose are key questions. Some work in this area has been undertaken, particularly by Yule et al in [195]. However, this tends to focus on ways to implement technology into the physicality of the moot so as to remove and replace face-to-face contact between members of mooting teams and the judges. Thus the authors evaluate the effectiveness of using the virtual online platform Second Life, a virtual classroom environment called Elluminate and videoconferencing to deliver moots.

“While there is general agreement as to the benefits to students of participating in mooting, the literature also points to a number of limitations inherent in the traditional model of mooting. [These] include overemphasis on appellate moots; limited opportunity to argue about the facts; the restriction on students being able to draft their own grounds of appeal; emphasis on oral rather than written submissions; lack of feedback; and lack of opportunity to develop an awareness of ethics and values. The use of technology might address these concerns and increase the opportunities to moot.” [195]

Although the above work identified benefits to using technology for facilitating moots, it also raised significant problems with the various approaches that were considered. There is a need for specialised information technology support when using virtual environments like Second Life because of their extreme abstraction

and complicated technical construction. Students felt that the effort in using these environments would not be worthwhile, for example. Simpler solutions like Elluminate resulted in a poorer development of advocacy skills because the students could not see the judges. They started to read their written submissions as a result rather than giving a spontaneous oral presentation which was simply guided by notes. Video conferencing resulted in a poor visual experience and required such a level of technical support to establish and maintain sessions that it inhibited the conduct of the mooting exercises to a substantial extent. Paliwala extensively reviews work in Computer Assisted Learning for lawyers and he concludes that:

"...the early pioneers of CAL avoided crude attempts at replacing people with technology. Any such ideas have always come and continue to come mainly from the people who control the money. During [platform] development we were asked by the funders "How many bums will this put on seats?" Of course resource saving is important otherwise money disappears down the money pit. Nevertheless, proper integration of technology to advance learning refashions personal contact teaching. Electronic [tools] are, in principle, enhancements of traditional learning resources and not substitutes for traditional teaching."
[145]

It seems, then, that the application of technology to facilitate moots in synchronous but remote collaborative environments is problematic. Much less work has been conducted to evaluate how the technology used in legal research prior to the moot session enhances or inhibits case preparation. It is in this area that this thesis makes a contribution. It is important to consider whether the construction, operating paradigms and user interfaces of legal information platforms are optimally designed for law students and trainee lawyers as they conduct research work. This thesis will address the question of whether computer-based research tools play an important part in promoting or inhibiting the development of professional skills and how any problems could be ameliorated in new platforms.

2.11 A boundary limitation: Other forms of applied training for lawyers

“Few law schools within the United Kingdom (UK) university sector have integrated clinics established as legal practices that offer live client work to the student body. Clinical legal education is becoming increasingly popular within the sector as it provides numerous advantages to the student cohort and establishes an opportunity for the students to gain important practical experience... the expansion and subsequent unbridling of the provision of a law clinic in the sector will provide the students with the skills necessary of graduates in the increasingly corporate, commercially motivated, UK university sector.” [126]

A law clinic is an organisation set up by a university law department or law school which provides free legal advice to members of the general public. Traditionally, legal clinics have served two aims: to provide free, or *pro bono*, access to legal advice for members of the community who would otherwise struggle to access the legal system; and to provide aspiring lawyers with hands-on opportunities to practice their craft. They allow the student to solve or address real legal problems whilst serving the tangible needs of members of the public. Clinics are perhaps the most realistic form of simulation that can be offered to students because, whilst the environment of a law clinic does not necessarily recreate the precise parameters of commercial practice, the issues that are dealt with, their ramifications and the skills required in order to give effective advice are entirely authentic. Law clinics provide for a high degree of reality in training whilst offering students the safeguard of supervision by qualified lawyers, who are usually faculty members at the host university.

In this context, it should be noted that the work presented in this thesis concentrates on a requirements analysis which is derived largely from the simulated environment of mooting. Other, potentially more advanced forms of simulation (and alternative frameworks under which students can address real legal scenarios) are not considered in depth. This is because, although law clinics are an increasingly common part of the educational landscape in the United States, penetration for the model in England (and the United Kingdom more generally) is limited at the moment. A design decision was therefore taken to concentrate on

mooting as it is the most established form of practical training experience available to law students in England. Almost every law department in the country will have a mooting society; far fewer will have an established law clinic through which students can practice the law in a supervised context.

At the same time, it should be noted that the interactions with solicitors which underpin the findings of this research as they relate to commercial legal practice include the requirements and feedback of several solicitor-advocates. A solicitor-advocate is a solicitor who also argues cases (a task usually reserved for a barrister), either in front of tribunals (as discussed at the end of Section 2.5) or in the lower courts. Thus there is a good reflection of the profession throughout this work and the boundary limitation which has been explained above should not be significant in skewing the results of the work as a whole.

2.12 Conclusion

The system of common law in England grew up over centuries, based initially on Anglo Saxon law, before being systematised and centralised. This allowed for more effective administration within the country and it streamlined the collection of taxes. The early development of a single system of law for the entire country culminated in Magna Carta in 1215. The common law system gives rise to a key principle that the law should be fixed and determined so as to remove the threat of the arbitrary use of power. The law should be accessible so that people know or can find out what it is and can plan their lives in accordance with the law. This key constitutional tenet is known as “the rule of law”. The idea that law should be accessible, clear and intelligible flows from this and it is important in the context of this thesis.

Interpretation of statute under either the textualist or intentionalist approach places a significant burden of linguistic analysis on lawyers and judges. The context of language use is of critical importance here. Access to legal language using computers must therefore reflect the importance of comparison, variation and context. The doctrine of *stare decisis*, or precedent, also serves to structure and link cases linguistically. It means that like cases should be decided alike. Precedent is a coercive and strong doctrine under English law. This means that development of the common law progresses through chains of linked authorities which apply legal principle to new scenarios. Any effective tool for legal research

must highlight linkages between cases and must recreate or preserve an idea of the weight of judicial findings from different levels of the court hierarchy.

Legal language is technical and syntactically complex. It comprises specialised legal terms with meanings that are specific to jurisprudence and also overloaded terms from everyday language. Some of these specialist terms are in Latin and need to be defined and understood much more carefully than a simple translation would suggest. There is thus a need to understand detailed meaning in everyday language and the domain-specific definitions which have been built up in the domain of law itself.

The traditional model of legal education, which persists from the late nineteenth century, is doctrinalist and teacher-led. There is relatively little emphasis on fostering and developing broad social skills like teamwork, collaboration and computer literacy. This is concerning in a context where legal information retrieval has moved almost entirely online. Efforts to change legal education have instead focused on increasing the diversity of a broad student body by providing new routes to qualify as a lawyer. That is not to say that there is no impetus to extend the methods used in legal education. Simulations of legal scenarios and environments have been practised for many years to a greater or lesser extent within the legal curricula of different law schools. Much of the modern work in this area looks at applying technology to produce situations where legal knowledge can be applied in realistic scenarios.

This problem-based learning strategy is exemplified strongly in the traditional discipline of mooting. Moots attempt to simulate the realities of presenting legal cases in court. Efforts to apply technology here centre on broadening participation by removing the need for face-to-face interaction so that remotely-situated students can participate. Results have been mixed, however, because of the new burdens and technical limitations which the technology itself introduces. It also seems to be a potentially dangerous idea to focus on broadening participation when location and immersion within the environment of a courtroom is one of the principal goals of the mooting exercise.



CHAPTER THREE

RELATED WORK

3.1 Introduction

This chapter discusses prior work which is related to the topic of the thesis as a whole. It is broken down into sections which relate together to form the background for the original research which is presented here. First, the major developments over the last four decades in electronic and computer-based legal information tools are summarised. The issue of a competitive duopoly between major information publishers and its effect on innovation are considered. Next, the results of several studies (predominantly from the United States) into how lawyers work with technology are discussed. A need for more information about legal working practice and computerisation in the United Kingdom under English law is identified. Work from the United States which has suggested a significant and growing skills deficit in the research abilities of law graduates and trainees is then related to the design of computer-based tools for information seeking. A key proposal in this thesis is that software tools which help to develop linguistic analysis skills in law students and early-stage lawyers can assist in improving levels of ability in and facility with legal research tasks. To this end, the domain of corpus linguistics and its potential to change the way that relevant information is identified from large bodies of unstructured text is considered next.

It is suggested that a move away from domain-specific, established result visualisations from corpus linguistics will be required to make tools accessible to lawyers and laypeople. Several existing visualisations are considered in this summary. The topic of data visualisation is then expanded to describe different interfaces that have been created specifically to work with legal data. The natural

interconnectedness of legal sources through precedent and statutory interpretation is highlighted as a key enabling factor in the creation of new visualisations for lawyers. Finally, the importance of collaboration as a skill for lawyers in both education and practice is discussed.

3.2 Computers and the law - information discovery platforms

“The great disadvantage of confining oneself to textbooks and lecture notes is that it means taking all one’s law at second hand. The law of England is contained in statutes and judicial decisions.” [191]

The above quote highlights an ideal that lawyers should be able to understand, manipulate and apply principles and language from the full text of prior case reports in different novel situations. This endeavour has been aided by the development of comprehensive digital archives of legal materials together with search interfaces through which relevant information can be found. From the 1970s onwards, companies like Thomson Reuters and Reed Elsevier began offering computerised information products to law students at university and to private law firms. These provide convenient access to large catalogues of legal case reports and items of legislation. This digitisation and search and retrieval effort first centred on legal materials under US jurisdiction before becoming available for English law in the United Kingdom. The venue for legal research was and continued to be the university library or the privately-held law libraries of legal practises. Law reports with judicial decisions from case law and statute books with records of legislation were previously only available in printed volumes through a small number of official publications to which the university or the law practice subscribed.

As discussed in Chapter 2, law reports prior to 1865 were published privately by a multitude of authors under their own names. These collections are called “nominate reports” for that reason. Altogether there were hundreds of different series of these private reports although many publications ran only for a short period of time. After the creation of the Incorporated Council for Law Reporting, case law publications became semi-official and were organised into a much smaller series of consistent publications. Old cases of significance from the nominate

reports were republished in the All England Law Reports, which were abbreviated in library classification systems to “All ER”. There is also one series of law reports for each division of the High Court: the Queen’s Bench division (abbreviated to “Q.B.”); the Chancery Division (abbreviated to “Ch.”); and the Family Division (abbreviated to “Fam.”). These series contain judgements at first instance in each division of the court and they also contain judgements on appeal to the Court of Appeal. If a case was taken further on appeal to the House of Lords, the decision would be reported in a separate series called “Appeal Cases” and abbreviated as “A.C.”.

Since 2001, the Incorporated Council of Law Reporters has itself sought to facilitate the move of case law from printed volumes to online databases. Any recent volume of case reports will feature paragraph numbers throughout the judgements. This was introduced in order to facilitate online discovery of specific information from cases and to enable a resolution of position and information between computerised and physically published collections. At the same time, a system of neutral citations for reports was devised and implemented. This sought to move citations for cases away from specific publications so that cases could be located and searched without reference to the old divisions between specific printed volumes.

As case law has been published reliably and has been systematised for online discovery and reference, statutory law was left behind somewhat. This is because there is a much more complicated scenario under which individual items of legislation are conceived and created by Parliament. Statutes are not arranged according to some rational plan which serves to classify and to subdivide them. The same subject or area of law may be subdivided between many different items of legislation. Statutes are also amended from time to time so that the state of the law often has to be ascertained by reading many different items of legislation side-by-side. Legislation is sometimes rationalised, however, in consolidating acts. These seek to bring together disparate elements of previous parliamentary decision making, but as Smith states [191], even consolidating statutes are unlikely to state the whole law on the subject with which they deal. *“The process of setting out both statute law and common law as a single, well-ordered body of law is called codification, but for various reasons English lawyers were historically hostile (or, at best, indifferent) to this.”* Nevertheless, the utility of digital tools in navigating the disjointed whole of different sources of law has long been appreciated.

3. RELATED WORK

“A researcher faced with interpreting language in an insurance contract can have Lexis find all other cases that quote the exact language in their text. A researcher can quickly assemble a collection of all cases in which the opinion is authored by a particular judge...by having Lexis search for all cases containing the judge’s name. Lexis thus frees the researcher from the constraints of formal indexing and permits every word, phrase, or number that appears in the text of a case or statute to be used as a retrieval key.” [168]

However, the challenges and expense associated with transcribing, digitising and reconciling databases of case law and legislation in an environment where the printed volume was dominant for years means that initial moves to computerise legal information retrieval were limited to and driven by a couple of large commercial companies in partnership with legal organisations. The Lexis system was the first digital repository of legal information. It initially prioritised making case reports and legislation available. Legal comments, professional literature and academic commentary followed somewhat later.

Lexis was first released in 1973 and was joined by competitor Westlaw two years later. These companies hired their own workers to transcribe, scan and digitise records of court proceedings, a significant enterprise which led to court cases about copyright infringement and unfair trade practises in its own right [197]. The scale of the information gathering exercise meant that the competitive duopoly between Lexis and Westlaw continued for a long time without challenge from other organisations or commercial companies. The ability of the organisations behind Lexis and Westlaw to short circuit the usual methods for collecting and creating case law reports of guaranteed quality has led to a massive growth in the amount of information which is available through these sources. Content completeness has tended to move the venue for legal research away from specialist law libraries with trained and experienced staff who can help to guide information seeking and search direction, however. [5] laments this side-effect of online search and retrieval because the influence of an important arbiter of relevance and source significance has been marginalised.

“The development of computer services in the legal information market has significant implications for law and legal practice. The use of computers to search full-text databases in legal research has been characterised as fast, objective, and flexible. Through online research, lawyers can complete

tasks that would have been extremely difficult or impossible with print sources. Computer-aided legal research has also changed the use and function of the library. Use of computers in legal research may also change the nature of precedents that might be consulted in a case. Computers may have also promoted use of persuasive authorities from other jurisdictions. Furthermore, computers have virtually unlimited storage capacity [and] allow rapid distribution of court opinions within hours or days.” [5]

The increasing uptake of digital legal information services brings many benefits, particularly in recent years with the move of Lexis and Westlaw to the online environment of the Internet and the World Wide Web. Convenience of access has been a significant factor here. Lawyers are now able to use information services anywhere where they have a computer and connectivity. They can search at any time which frees them from the environment of the law library and allows for “just in time” access to reliable and relevant legal information. The move to digital resources is not without its own problems and implications for lawyers, however. Commercial databases for legal research tend to be large and monolithic. An increasing drive to provide “digital completeness” in information coverage leads to content integration. As a result, search interfaces become more complex in order to provide access to the breadth of data which has been published. The state of user interfaces for legal search was described by a senior legal practitioner during the course of this research as “catastrophically poor”.

Law schools only run courses to provide basic familiarity with search functions thanks to the complexity of user interfaces [140]. System designers expose broad swathes of data by showing many results for a given search query. Many professionals and commentators find this to be confusing and dispiriting. Coverage has come at the expense of system usability. Another problem is that the huge range of legal information sources which are available online is not necessarily expertly or effectively moderated [59]. This is said to risk a degradation in the quality, relevancy and completeness of legal research. Training in information technology which is fit for the purpose of online research from wide-ranging legal and non-legal sources is now an important issue for legal education providers. Some commentators suggest that law school curricula and professional development schemes have not caught up with the challenges of using digital information.

“The point is of course, that the basic functionality of most information

3. RELATED WORK

retrieval systems hardly presents problems to most users. But more advanced functions - such as using automatically generated cross-links to find relevant legal comments for certain legislation, or to use a notification function in such a way that only relevant new documents will be shown, or to add (parts of) retrieved documents to a digital dossier shared with colleagues - require additional study and practice, the time needed for which is often not invested.” [140]

Recent government-funded projects to digitise case law for open access in the United States (under legislation called Public Access to Court Electronic Records, or PACER) have resulted in the creation of new tools for online legal research. Casetext [27] provides full-text search and retrieval of case law which is socially augmented with user annotations. User-generated content is amalgamated and mined with machine learning algorithms. Ravel Law [152] adds visualisation to full text search by producing interactive heatmaps of text hits across cases. Knomos [97] employs open legal data to generate network diagrams of the connections between legal sources.

There has been some diversification of online legal research tools which cover English law as well. This has been less positively disruptive than in the United States partly because of the limited sources of open legal data that are available to technology companies here. This barrier results directly in a less diverse ecosystem of legal information products for the English lawyer. Many court cases are recorded and transcribed by private companies in England which have been appointed by the Ministry of Justice and HM Courts and Tribunals Service. The revenue streams of these companies rely upon state funding and also upon licensing fees from third-party information publishers. This publishing model necessarily limits the desire of the transcribers to become involved in open law initiatives.

“The second path, which I favour, I would refer to as the open access “ultra” model. Under this model, the government would bring all transcription in house and the transcripts would be made available online by the government in a form that is suitable for republication, reuse and data analysis (much like primary legislation on legislation.gov.uk).” [81]

That being said, JustCite is a British company which uses licensed content from

the major publishers of legal information to produce visual precedent maps. A precedent map is a radial diagram showing cases cited in an authority and how each citation was treated. JustCite relies heavily on expert curation under which a team of legally-trained staff establish the important content from case law reports and construct the precedent map visualisation so that it demonstrates the relationship between a case of interest and other significant authorities [46]. The Supreme Court has established a web site where their decisions are published for public access within hours of them being delivered.

Another state-run resource for English law is the governmental legislation archive. This collection includes the full text of all statutes that are currently in force in the United Kingdom. It offers various search facets - like act title, act date and a keyword search - which enable effective discovery. Crucially, the portal also exposes statutes through a pre-defined ontology of subject areas. The user can therefore find all acts which concern a particular topic with ease.

Finally, it is worth mentioning the Practical Law service which is now offered by Thomson Reuters. This commercial archive provides legal form templates, accounts of the state of the law by subject, question and answer scripts and other pre-populated documents which are uploaded by lawyers with different specialisms. The goal of the system is to help and guide other lawyers in their research and preparation activities.

3.3 How lawyers work with technology

“The work of lawyers is highly varied. It can involve criminal matters, corporate matters, regulatory issues or private disputes of various types. Some lawyers handle a wide range of types of matters although increasingly lawyers have tended to specialise. Lawyers work in a wide range of settings, from firms with literally thousands of lawyers to solo practises to government and corporate offices. While law school provides the foundation for persons to enter the legal profession, the actual practice of law is something that is ultimately learned by doing...Essentially, during the early years of practice, the new lawyer learns the craft of practice.” [103]

Herbert Kritzer follows this introduction with a collection of essays about how lawyers conduct their work in [103]. These essays are based upon the findings of a

series of research projects for different institutions over a thirty-five year period. It represents the most up-to-date and comprehensive account of legal activity which is presently available. Kritzer points out that arriving at an understanding of how lawyers work in their day-to-day activities is time consuming, sensitive and expensive, especially when data is obtained through observation. He starts with a thesis that direct observation of people at work is the best way to collect accurate work activity data. It enables detailed questioning to elucidate why activities are undertaken and it avoids the subjectivity of questionnaires and surveys which are delivered after the fact. This thesis uses a similar approach to find out how mooted students and then solicitors work together in Chapter 4.

However, there is no focus here on the use of digital devices and electronic information platforms. In fact, many of the research projects which give rise to the results in the book predate the availability of electronic legal information tools. Kritzer notes at the outset that much of the data was collected for a United States government project called the Civil Litigation Research Project in the late 1970s and early 1980s. His datasets therefore cover topics like the operation of lawyer-client privilege, the way lawyers work in a “no win, no fee” environment, the commodification of certain areas of the law like insurance services and the differences between legal culture in different localities.

The 2013 book *Tomorrow's Lawyers* examines the role of technology, the Internet and digital working practice on legal practitioners [171]. The author attempts to predict the effect that technology will have on the legal profession in the future and how the practice of law will change as a result. The influence of technology and its implications for lawyers is only one of three drivers for change in the profession which are identified and discussed. Susskind sees change in the legal profession emanating from: a public expectation that lawyers should provide a broader range of services in return for reduced fees; liberalisation so that it will no longer be universal that the people who provide legal services are qualified lawyers; and information technology centring heavily around artificial intelligence both to serve lawyers themselves and to replace them. There is no particular focus on case preparation or information discovery here, however.

An early analysis of how lawyers used digital information retrieval is presented by Blair and Maron in [17]. This is a technical paper which sought to discover how accurate and comprehensive legal search engines were, in terms of precision and recall, as compared to the level of confidence that users had in their ability

to discover all documents that were relevant to a search query. Values for precision and recall were established using statistical sampling methods and blind evaluation procedures. The results show that the search engine was retrieving less than 2% of relevant documents in response to queries whilst users believed that they were seeing 75% of the critical material that they would expect from experience. Although this work is now over thirty years old, and computers and search technologies have evolved greatly in that time, the conclusion of the authors that search systems should be based upon the full text of a document collection rather than on post-facto enrichment of the data is directly equivalent to the approach taken in this thesis.

“Data retrieval by subject content...[eliminates] the richness and flexibility of natural language [which] have a significant influence on the conduct of an inquirer’s search. The inquirer must describe his information need using subject descriptors which have been assigned to documents...The indexer must choose appropriate terms to describe the information content...But there are no clear and precise rules which an indexer can follow to select appropriate subject terms.” [17]. This conclusion is supported through later research by Burkhard Schafer, in which he demonstrates the lack of scalability in manual annotation of legal sources by stating that: *“proponents of knowledge engineering in legal retrieval observe that landmark cases are not necessarily discernible from analysis of text and important future legal concepts may not be mentioned at all. However, the knowledge engineering approach suffers from the need for highly specific annotation of legal issues, concepts and factors, and detailed knowledge bases encoding the ways in which they interact.” [130]*

“Information-seeking is an important part of lawyers’ work and unlike many other professions, the legal profession has access to many dedicated electronic resources. Despite access to these resources, lawyers often find legal information-seeking difficult, making them interesting to study. Much of the problem might lie with the fact that digital law libraries have traditionally been regarded as difficult to use.” [123]

From the position taken in the above quote, Makri et al [123] summarise the findings of a small number of user-centred studies of how lawyers work with computer-based information discovery tools. They conclude that existing platforms are not optimally designed because lawyers have difficulty in formulating appropriate search terms. Lawyers also find it difficult to understand and to use

3. RELATED WORK

the special and individual search features of different resources within the same platform. Some of the studies point to the fact that there is no way for users to know when they have exhausted all possible search avenues in an attempt to find relevant materials for their legal research task. It is also suggested that long exposure to and experience of existing platforms, particularly by law students, does not tend to result in reduced error rates or an ability to more quickly find relevant information. Many advanced features and commands on the systems were never used.

Some existing design-focused studies of how lawyers work with technology centre on the creation of systems and tools for information use and re-use rather than on information seeking. A study by Blomberg et al [19] involved the creation of a digital filing cabinet platform for a legal company in the United States. The solution was designed to provide a bridge between physical and digital documents so that scanned versions of important information could be stored centrally on computer. The system allowed this archive to be searched in a manner that was dictated by watching how the lawyers in the firm worked and what their existing information-seeking behaviour was.

Another important study by Marshall et al [125] used observation data from a mooted preparation exercise at a university in the United States to design an e-book reader. The reader gave wireless access to important information sources and allowed students to annotate documents with their own notes and other content. Komlodi and Soergel [101] observed lawyers working to determine how memory from prior experience and electronic search histories facilitated or blocked effective information re-use. This was taken further a few years later to examine how electronic search histories could be used as a collaborative resource for facilitating legal research [100].

“[S]earch history creation and use are naturally occurring information behaviours and are accomplished regardless of IT support. [People employ] manual work-arounds in the absence of adequate history tools, and, where support did exist, [there are] limitations of that support. Our results confirm previous findings that contemporary search history tools are not living up to their full potential. They are too narrowly focused on supporting single users completing specific information tasks.” [100]

Kuhlthau and Tama [104] undertook a series of structured interviews with a group of practising lawyers in the United States. The idea was to better understand how lawyers acquire and use information in their work. The study evaluated how the participants approached information-seeking tasks, how the tasks were broken down into stages and how outputs from these stages were subsequently unified for knowledge synthesis. The authors identified a lack of computer-based tools to store information in a form that was accessible to groups of lawyers.

The collaborative nature of legal research was again addressed by Jones [89] through a series of contextual inquiry exercises at a legal aid clinic in the United States. The sessions were recorded in video and audio and transcripts were taken for further investigation, which is the same approach adopted in Chapter 4 of this thesis. Jones's findings highlighted the social nature of legal information seeking and the relative lack of platforms which enabled group working through information re-use.

"Law is a knowledge-based profession and in its core "legal practice" is about providing specialised knowledge and services in a variety of ways to a variety of clients. This knowledge, or intellectual capital - the law firm's aggregated experience or collective wisdom, applied to delivering knowledge-based services - is one of the most important assets of a law firm. Yet traditionally, many firms have taken an ad hoc approach to managing this asset, resulting in work duplication, inconsistent work practises and loss of important organisational knowledge when lawyers retire or leave the firm." [51]

As quoted above, du Plessis and du Toit [51] examine the effect of information technology and digital legal libraries on legal research. They attempt to identify the skills which practising lawyers have, as experienced workers in print-based archives, that can be transferred online, and which skills they need with computers that must be newly acquired. This effort focuses on guidelines for the implementation of knowledge management systems in South African legal firms. The authors point to the problem that a jigsaw of tools exists along with a multitude of methods for storing digital data and for then managing and leveraging it. Their study concludes that the broad range of different specialist and general tools which is available to the modern lawyer inevitably leads to the creation and exacerbation of skills deficits.

Margaret Wilkinson [190] cautions against an absolute conflation between legal research and information seeking. She points out that information seeking is a broad activity within law firms which encompasses much more than research and case preparation. Her findings indicate that lawyers were more concerned with problems in the administration and running of their practises, and in finding the requisite information to do this efficiently. This helps to explain why many research efforts and products in the legal sphere centre around knowledge management and digital archiving of case files and documents pertaining to the lawyer-client relationship. Interestingly, Wilkinson highlights results from a series of interviews with practising lawyers which expose a preference for informal, general information search tools over specialist platforms. That is a finding which will be probed more deeply in the rest of this thesis.

3.4 The research skills deficit

“Law schools are confronting a sea change in their educational responsibilities as they contend with calls to instil skills training in addition to teaching doctrine and analysis. In addition, ever-growing waves of information are overwhelming law students, eroding their research skills, and weakening their ability to learn legal analysis.” [183]

“[Lawyers need to be] able to find exactly what they need with less floundering and more precision. When time is a valuable commodity (either because it is being billed or because there just isn’t enough of it to get everything done), hours saved from fruitless, inefficient searching...would pay off.” [10]

The importance of skills training, and particularly the ability to effectively conduct legal research, which is reflected above by both Valentine and Barkan, must be evidenced in the composition and characteristics of law school and university education programs for lawyers. Almost all institutions have courses that seek to develop and refine analytical abilities for research and writing in students and trainees. These elements of the curriculum are sometimes compulsory but are widely available as elective classes. The ability to research competently and to write clearly are pre-requisites for a practising lawyer [57]. To effectively serve clients, the search efforts of the legal practitioner must be thorough and complete. Such is the burden of cases and ever-increasing numbers of reported judgements

on courts that judges have stated that they cannot be responsible for highlighting unidentified issues or legal arguments - see *R v Boardman* [2015] EWCA Crim 175.

The process of skills training in universities and law schools is complicated by the proliferation of new areas of competency which the adoption of computing technology has created. Where once there was a clear focus on the importance of mooting, on public speaking and on the creation of effective written documents, there now exists a huge range of skill sets that are candidates for teaching and development in aspiring lawyers.

“Notwithstanding the ubiquitous presence of computers and the Internet at most American law schools, little has been done to expose future attorneys to the role that information technology will play in their professional lives...The range of technologies has exploded - information sources and techniques are proliferating - but the standard means to keep track of and filter information have not kept pace.” [79].

In that context, it is perhaps unsurprising that focus on the core competency of legal research has been diluted by the introduction of other professional skills which must also be taught. Valentine [183] highlights two recent studies from the United States which both concluded that law schools are broadly failing to teach fundamental professional skills that are required for the competent and ethical practice of law. Callister [25] devotes an entire section of his paper on the role of librarians in teaching legal research to comments from the United States bar association, practising lawyers and others who have perceived declining competencies of legal research in new law graduates over a relatively long period of time. More importantly, however, criticism about the alleged decline in legal research skills comes from formal studies which expressly seek to evaluate such abilities:

“There is a growing awareness among law librarians and practising attorneys that the research skills of law students and recent law school graduates are painfully inadequate and are perhaps becoming increasingly so. The survey confirms the perception that most summer clerks and first-year associates are unable effectively and efficiently to research issues that appear routinely in cases handled by middle-sized and large law firms.” [119]

A 2013 survey from the United States reported by Susan Mart [127] found that, despite declining abilities, lawyers spend a significant portion of their time on legal research. From a total of six hundred respondents, half spend approximately 15% of their time researching the law whilst 10% spend half their total working time on this activity. The survey demonstrates a correlation between the number of years that a lawyer has practised and the amount of time they spend on legal research. Legal research is delegated to those members of firms who are recently qualified. Paradoxically, this means that staff with the lowest levels of research competency from their law school backgrounds are tasked with the greatest proportion of research work. In the same study, 40% of two hundred senior lawyers said that recent law graduates perform cost-effective research “poorly” or “unacceptably”. Within that result group, almost half of respondents said that new entrants to the profession were able to construct and implement effective research plans only “poorly” or “moderately well”.

“As long as state-of-the-technique requires multiple media, and especially after it requires interactive video, electronic imaging, artificial intelligence in law, and as-yet unknown new media, the problem of integrating all that needs to be taught, learned, practised, and refined about legal research into the law school curriculum will be one of legal education’s most difficult challenges.”

[52]

The history of poor satisfaction with the research skills of trainee and recently-qualified lawyers from the United States is instructive. However, there has been little formal research to establish whether the problem is replicated in the United Kingdom. Even more significantly, the role of technology in ameliorating or exacerbating this deficit warrants further investigation. There is a tendency in the available literature to blame computers and online information platforms for skills shortages. However, this has not been formally addressed either to find out the truth of the supposition or to expose why computer-based products are problematic. The rest of this thesis, and particularly the contextual enquiry in Chapter 4, will address this research need through the lens of user-focused work studies related both to law students and to practising lawyers in the UK.

3.5 The growing role of corpus linguistics

A fundamental characteristic of the practice of law is the production and consumption of documents. As a result, the creation of tools for handling and searching within written materials which apply effective and principled methods for exploring and navigating text presents key challenges here. The scientific study of language and its structure through techniques from linguistics can be useful when seeking to better understand and manage such sources. Today, computational linguistics in general and the field of corpus linguistics in particular afford powerful approaches to tackle many of the challenges about information management and information overload that are faced by lawyers.

“The principles of corpus linguistics have been around for almost a century. Lexicographers, or dictionary makers, have been collecting examples of language in use to help accurately define words since at least the late 19th Century. Before computers, these examples of language were essentially collected on small slips of paper and organised in pigeon holes. The advent of computers led to the creation of what we consider to be modern-day corpora.”

[13]

Corpus linguistics is the discipline of studying language in use through the evaluation and interrogation of corpora using computers. The new science initially revolutionised lexicography, as reflected in the quote above. A corpus is a large, principled collection of naturally-occurring examples of language which is stored electronically. A key idea behind modern-day corpora which are considered to be well-formed is that they should be representative of the type of language which they exemplify [147]. This means that a corpus should be of sufficient size and scope to capture accurate and repeated details of how language is used by different sections of society or in different idioms and genres. Corpora are usually composed of complete written documents so that each constituent part of the collection is structured reliably and appears within a valid context, although some collections use fragments of text which are brought together, particularly where the field of study is spoken language.

The design and composition of corpora is an endeavour that needs to be approached carefully on the basis of a proven philosophy because conclusions about language use which are drawn from them must be empirically accurate. The

first digital corpus of language which could be interrogated by computer was called the Brown corpus and it was released in 1961 [175]. This collection featured one million words which was composed of five hundred written documents that were originally published in the United States during the year that the corpus was released. Modern corpora can now include hundreds of millions of words of text. Advances in computing technology and the efficiency of search and retrieval algorithms is starting to facilitate the practical interrogation of collections that are billions of words in size.

One of the most prominent and influential scholars in the field of corpus linguistics was John Sinclair, Professor of Modern English language at the University of Birmingham between 1965 and 2000. Sinclair proposed and pioneered many of the core principles of corpus design, practical composition and methods of interrogation which are now considered fundamental to the discipline. His key idea was that a single word in itself does not carry or communicate meaning. A word does not have atomic meaning.

Word meaning derives, instead, from contexts of usage. As the linguist J.R. Firth commented, *"You know a word by the company it keeps"* [55], and as the highly influential philosopher, Ludwig Wittgenstein also pointed out: *"For a large class of cases - though not for all - in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language"* [194]. As Sinclair points out, we convey meaning by using collections of words in a sequence. This led to a proposition called the "idiom principle". The idiom principle holds that we do not select words in an open choice environment when we talk or write. Instead, we select pre-prepared or commonly used sequences of words which we know to be valid from experience and bolt them together in new writing or in new utterances so that we can be understood.

Corpus interrogation software allows for corpora to be searched so that key sequences of words may be identified in response to queries. Corpora can show us the context in which individual words are used and it is these contexts rather than the word of interest itself which impart meaning and are of significance. Sinclair proposed that the collocation, or two words which occur together, was the smallest indivisible unit of meaning. Collocations are in fact words which appear together in a corpus more frequently than chance itself would predict [12]. This means that they are candidates for pre-selection by language speakers and writers and that they form the fundamental building blocks under which the

idiom principle operates. According to Goldfarb, “Sinclair recognised that this [idea] would raise problems which are not likely to yield to anything less imposing than a very large computer.” [63].

“I knew it would be necessary to modify the traditional concept of the word. But the idea now dominating my work [is] that the unit of meaning is rather a phrasal unit than a word. Once we accept that words can be co-selected, not chosen always one at a time, then there is no longer a problem with “dark night”. “Night” does not distinguish one of the meanings of “dark” and “dark” does not distinguish one of the meanings of “night”.” [165]

The corpus-based approach to linguistic analysis may be principled and based on a clear philosophy about how people use language, but it is not restrictive. Any well-designed collection of text from any era and any genre can be codified into a valid digital corpus for interrogation by computer. The corpus approach to linguistic analysis holds only that: the endeavour must be empirical, analysing the actual patterns of language use in natural texts; that it utilises a sufficiently large and principled collection of texts as the basis for analysis; that it makes extensive use of computers for analysis; and that it depends both on quantitative and qualitative analytical techniques [15]. In this light, it is not surprising that people have built corpora of legal texts for analysis by computer.

One of the earliest such collections is the *Old Bailey Corpus* of legal court reports from the English legal system. This corpus is based upon digitised transcripts of the proceedings at London’s Central Criminal Court, which is known as the Old Bailey, from between the years 1674 and 1834. The collection totals some fifty two million words of transcribed spoken English and it has been annotated with social and biographical detail from the court records about individual speakers, where that information is available [84].

Another legal corpus of English law sources is The British Law Report Corpus (BLaRC) [22]. It is a small collection of eight million words of text from case law which covers the period 2008 to 2010. Another collection which is relevant here is the Corpus of Historical English Law Reports (CHELR) [34]. This is a very small dataset of less than 500,000 words which includes some case reports for the period 1535 to 1999. Any freely-available corpus of English law sources is useful but the small scale of these publications, together with the fact that they are static, make

them of limited utility as tools for enabling legal research. The creation of modern legal corpora depends heavily on the availability of open access legal information.

Although the British and Irish Legal Information Institute was established nearly twenty years ago as a charitable organisation to provide open access to legal judgments under UK and European law, the organisation has not yet secured equality of access to legal information with the major commercial publishers. As Philip Leith states, “...what is surprising is that the open access model in legal information still appears to have opponents within the group who control access to judgments, and consequently BAILII is never completely successful in getting the judgments it requires to satisfy the needs of its users. For example, in a recent attempt to secure eight judgments that a government body wanted posted on BAILII so that they could be included in their training coursework materials, only one was obtained despite the fact that six of the remaining seven were obtained by Westlaw and Lexis.” [111]

“[F]our issues should be addressed before turning to corpus linguistics as the most efficacious tool in statutory interpretation. First, the legal issue before the court must be about the distribution of linguistic facts. Surely, separating the “ordinary” sense of an expression from outlying ones meets this criterion...Second, along these same lines, the court must decide, as a legal matter, what makes an interpretation “ordinary”...Third, if one wishes to search a corpus to glean the ordinary meaning of a term, one must decide, in advance, what to search...Fourth, there are two very different reasons for a particular meaning to present a weak showing in a corpus search. In some instances, it is possible, but awkward, to use a particular expression to describe an event or a set of circumstances...In other cases, a particular usage may be absent from a corpus not because speakers are uncomfortable using the expression in that way, but because it reflects relevant circumstances that do not often arise.” [166]

In recent years, the utility of large-scale and well-designed corpora to provide evidence about the meaning of language has started to have an impact directly on the justice system. Corpus evidence about what particular language usually means - to “the man on the Clapham omnibus” as defined under English law (to mean a reasonable and ordinary person) - can be a valuable asset in assisting judges to interpret statute, to understand and apply contractual terms and to gauge the intention of parties to a legal dispute. This reflects a growing trend to move away

from the idea of the judge as a socially-connected arbiter of meaning and intention to a vessel through which empirical analysis of what people write and say can be determined [185].

“[L]exical semantics and other aspects of language are integral to legal interpretation. As such, inaccurate judicial assertions about language, which various scholars have catalogued, sometimes result in interpretations that might not have been selected absent incorrect understandings of language.” [66].

There is widespread agreement about the primacy of the “ordinary meaning” rule in legal interpretation but a debate has started about who is best placed to decide what is ordinary and how this decision should be researched and informed. One problem here is that the pool of available corpus linguists is finite and the process of detecting meaning requires expertise and experience. The tools that corpus linguists use to interrogate collections of text are not designed to be familiar to lawyers or judges. Some scholars address this concern by advocating, or at least examining, the broader employment of empirical linguists as expert witnesses [36]. This idea is fraught with problems about how the linguist understands legal issues and proceedings, for example, and it may appear to be a largely impractical avenue. Until such time as corpus analysis becomes commonplace in the legal profession, and legal experts are given access to resources for linguistic analysis which they can work with, expert witnesses may be one way of addressing concerns about inadequate judicial appreciation of ordinary meaning.

“When we speak of ordinary meaning we are asking an empirical question - about the sense of a word or phrase that is most likely implicated in a given linguistic context. Linguists have developed computer-aided means of answering such questions. We propose to import those methods into the law’s methodology of statutory interpretation.” [110]

The idea of employing corpus linguists to provide evidence about the ordinary meaning of text in a legal scenario tends to move interpretation towards a newly-defined approach which is related to traditional textualism. Corpus linguistics claims that the definition of meaning is an empirical question which can be answered most effectively by how often a given term is used in a particular

manner. The most frequent context for usage which can be found empirically from a well-constructed corpus therefore serves to reframe the definition of “ordinary meaning” [78]. However, courts do not simply define ordinary meaning as an empirical question. Sometimes judges use “ordinary meaning” to refer to whether a meaning is permissible under the law, sometimes to question whether a meaning is obvious, and sometimes to refer to the meaning that the hypothetical reasonable person would give to the statutory language. Thus, the application of a rigorous, empirical approach to linguistic definition in the law through the interrogation of corpora brings its own problems and is itself a contentious development.

“To make good use of corpus resources a teacher needs a modest orientation to the routines involved in retrieving information from the corpus and - most importantly - training and experience in how to evaluate that information. It is this second point that has caused much controversy, because a corpus is not a simple object, and it is just as easy to derive non-sensical conclusions from the evidence as insightful ones.” [163]

This pivotal role for the corpus linguist and the idea of ordinary meaning in the law does little to address the possibility of corpus tools which lawyers and judges can themselves use. Another avenue which provides interesting possibilities is the creation of corpora and interrogation tools that concern and that examine legal language and which are designed to be used by practitioners themselves. It is a key tenet of this thesis that legal language very often has domain-specific meanings which are separate, different and more developed than the ordinary meaning of the terms involved would convey. Thus it is possible to suggest that corpora used in the legal environment should also extend to defined collections of legal language.

Part of the problem here, certainly under English law, is the difficulty of obtaining large collections of legal documents, case reports and legislation without entering into expensive and restrictive licensing agreements with established legal publishers. Another issue is that corpus tools are themselves specialised and difficult for laypeople in the domain to use reliably. In the United States, Brigham Young University Law School has recently released an online corpus interrogation system to allow judges to explore language and meaning. This tool can also make use of a corpus of decisions from the Supreme Court [2].

From the standpoint of legal education - with an emphasis on education - there have been fairly widespread efforts to integrate corpus evidence into general language curricula. As Sinclair notes, the chorus of complaint about the principle of using corpus data in education has largely subsided in the face of practical results. This sets the scene for corpus tools to be used not only in legal practice but in training and education as well. In 2004, Sinclair himself consulted on a project to introduce a corpus search engine into English language education in Scottish schools. Feedback on a prototype of the system was mixed, however, because teachers found it difficult to integrate the search engine into their daily practice.

Ultimately, the PhraseBox product was never released. The issue of integrating corpus interrogation software into the work contexts of professionals is an important result of that project, however. It is suggested that the BYU tool for working with legal corpora, whilst interesting and useful to researchers, may not meet its full potential because it represents yet another piece of the jigsaw of software tools that lawyers and judges are increasingly expected to be proficient in and to use. PhraseBox demonstrates that, contrary to Sinclair's assertion above, the learning curve for traditional corpus tools is far from modest. Success may be found, however, in a development project which is driven by user-centred design based upon an evaluation of how lawyers work. This thesis pursues that idea on those terms.

3.6 Corpus interfaces: KWIC and other visualisations

"Within corpus studies, form-oriented language concordancing, in particular in the shape of KWIC (Keyword in context) concordances, has received most attention...This type of concordance is instantly recognisable. The rows of individual concordances combine to produce a semi-tabular format with a single central column identified by automatically-created alignments, bold type, colour and gaps, allowing users to perceive patterns in wordings and to relate them to their co-texts." [8]

Several corpus management and interrogation systems exist through which users can encode text in a suitable format and then interrogate the resulting corpus for information about language. The management software usually provides both

the tools and formats for corpus creation from plain text and the user interface through which results are obtained and displayed. There are no formal data formats for creating and indexing corpora but the popularity and prevalence of certain products leads to a degree of de-facto standardisation.

Some examples of commercial, application-based corpus interrogators include WordSmith [114] and ParaConc [11]. These are both long-established products which have attained a degree of standardisation within the corpus linguistics community. Free corpus management applications include AntConc [4] and CasualConc [87], while the CorpusExplorer [35] software is open source. Some concordancers are tied to specific corpora, such as the online interrogation interface called BNCWeb [174], which enables search and retrieval on the British National Corpus. Examples of online corpus management and interrogation products include CQPWeb [107] and SketchEngine [94], the latter having a related version called NoSketchEngine [128] which is free to use and open source. The Sketch products include a complete workflow for building and preparing corpora together with a concordancing user interface called Bonito, which allows users to work with the corpora that they create. A full list of corpus managers and user interfaces for corpus linguistics which are currently available is located at [33].

The traditional layout of search results in user interfaces for corpus evaluation is *KeyWord In Context* (KWIC). Almost all corpus managers implement the KWIC user interface design in some form. KWIC is a textual display of lines of language arranged vertically around a common or “lemma” node, where a “lemma” is a root wordform which returns search results for all forms of that root. The word form under examination - the original word or phrase that is being searched for - appears in the centre of each line, with extra space on either side of it. The length of the context around each node can usually be specified and altered for different purposes, but a context length of sixty characters to the left and sixty characters to the right is a normal default. Thus the user is presented by a group of partial sentences which are arranged vertically down the screen and are centred on the search query node [162].

Presenting results through KeyWord In Context is effective for a trained linguist but it makes corpus tools inaccessible and confusing for experts from different domains [77]. A key outcome from early testing of the PhraseBox prototype in Scottish schools was that teachers in secondary education found the paradigm difficult to understand. Another problem is that KWIC was designed to be

efficient with both horizontal and vertical space on screen and on the printed page. However, this means that there are limited opportunities for augmenting the display with associated, meta-textual information about the sources from which results are derived.

The annotation of the corpus with descriptive data was held to be undesirable unless that information was germane to linguistic analysis of the text itself. Therefore, segmentation of the corpus with genre descriptors and keywords was to be discouraged because these were sensible only to a particular corpus user or group of users. There was also a problem of subjectivity. Descriptions of items within a corpus might be appropriate for one interrogation purpose but they might be misleading or inaccurate for others. The corpus in PhraseBox, for example, was encoded with part-of-speech information derived from an English language tag library because this data was finite and would not change according to different usage scenarios. However, genre separation was accomplished by encoding multiple corpora, one for each type of linguistic resource, rather than by annotating one large corpus with topic and stylistic descriptions.

Another problem with the KWIC paradigm is that it is not very efficient at demonstrating linguistic variation around the search node. The vertical arrangement of lines enables analysis at different token positions but the nature of a corpus means that there will be many similar contexts for a particular query. If we take the phrase “cup of tea”, the two central vernacular meanings in English are usually “I would like a cup of tea” (the person wants a hot drink) and “it is not my cup of tea” (the person does not like something). Contexts for both of these senses of the phrase will be dominant, but peripheral variation within and between these senses is difficult to isolate.

3.6.1 Alternatives and additions to the KWIC paradigm

In order both to broaden the appeal of corpus tools and to better highlight variations of language around the search node, different visualisations can either partly or completely replace KWIC [41]. Most of these alternatives are hierarchical and use tree layouts. The user can explore different sentences by iteratively selecting word components and viewing the choices available to them once a selection has been made. *WordTree* [186] is a single-sided implementation of this approach, which means that either the left or the right context of a search hit can be explored at any time. *DoubleTree* [40] allows for exploration on either

side of a query node at the same time. These types of visualisation have been built into complete user interfaces, like the *Wordgraph* system [156]. Some new display paradigms utilise frequency information which denotes relevance in a visualisation like *Corpus Clouds* [42].

Corpus Clouds is a replacement for small contexts around a search node, such as the environment which is shown when collocations of a word are queried. Instead of the truncated KWIC result set that is usually provided, corpus clouds float words around the search hit. The size of the “bubble” in which the collocation is enclosed indicates its relative frequency in the result set. The proximity of each bubble to the central node is sometimes also used to denote the relative position of the collocate to the search token. Another approach to looking at collocation is presented by Kilgarrieff and Tugwell [95]. They propose the “word sketch”, which uses statistical salience calculations and part-of-speech data to tell an overall story about how a particular search word is treated in its common environments [94].

3.7 Visualising the law

“Law can be made more comprehensible if it is made more visual. This means illustrating cases - putting the human situations back into the legal opinions - creating flowcharts out of rules - and thinking about how we can convert complicated text into clear, digestible, graphic presentations.” [69]

There has been work in recent years to visualise legal data for better knowledge synthesis. Activities here can be separated into two different categories: using visuals to make law more accessible to the general public in publications like posters and infographics; and tools which facilitate understanding relationships between legal data which are aimed at computer scientists and researchers. The Open Law Lab [69] is an initiative which produces visual designs that depict the main points of complex laws in a way that laypeople can understand. *SketchLex* also creates accessible legal infographics aimed at the general public [105]. Helena Happio runs *legaldesignjam.com*, inviting contributors to redraft complex contracts clearly using simple language and visual design elements [70].

These approaches have been evaluated by government as potential new formats for publishing legislation [14]. *Kohvolit* is a company which creates interactive displays of the progression of bills through the parliamentary process [99].

Similar endeavours for French legislative instruments and visualisations of their development over time are available from [106] and [184]. The University of Michigan has also created network-based interactive maps from a range of legal datasets in the US [91].

Curtotti [43] provides a good summary of open access legal information repositories aimed at the general public which focus to some extent on visualisation. Many of the approaches in what is an emerging field of study have been formalised as communication guidelines which seek to move law out of its “*text-orientated universe*” [24]. However, the *status quo* in online legal research tools still relies on two products - *LexisLibrary* and *Thomson Westlaw*. Both products improve comprehension to some extent through visualisation. *Lexis* has introduced a timeline which charts the progression of cases through the court system. Search term maps also show how common query words are in case reports and how hits are distributed. In general, however, both products are still textual.

On the other hand, Ravel Law [152] visually represents the most important cases for a given search query as a network of connected nodes. Edges from the root node lead out to subsequent cases that have used the same language or have cited the same root case. The size of the hub for each connection reflects the relative number of cases that concern the given topic or that cite the given case. The frequency with which courts cite a particular case can therefore be used as a citation index, so that users can gain some understanding of the importance of each authority in their field of interest.

Aris [6] proposes a system for organising and displaying information about court cases. This is focused on case distribution through facets like court level and jurisdiction. The authors call the networks which are built up from this data “*semantic substrates*”. The purpose here is really to allow teams of legal experts to analyse the relationships between court cases with the assistance of computer scientists. A prototype tool for designing substrates is proposed. This approach is useful for identifying clusters of legal activity based upon different criteria.

The work does not translate well into a general-purpose visualisation tool for lawyers and legal students, however. Semantic substrates are complicated and there is a significant learning curve associated with understanding what the diagrams are trying to represent. They also do not allow for full-text access to the important parts of case reports and legislation which is represented in the

3. RELATED WORK

substrates. Perhaps most importantly, however, a significant amount of expertise and experience is required in order to build different substrates in the first place.

Branting [21] proposes a model for codifying connected legal precedents using a reduction graph. This allows for the key facts of a case to be delineated together with the decision of the judge. In between these fundamental items of information about different precedents, reasoning can be broken down into distinct elements and subsets so that the relationship - called an inference path - between the facts and the important elements of the judgement can be clearly connected and understood. The result is a directed graph which associates key facts with issues of law and the portions of the judgement in which each of these facts and issues is considered.

“When a citizen tries to understand how a given issue is legally disciplined, when a legal professional tries to see how a specific area of a legal system evolves over time, their attention cannot be limited to a single source of law. Specifically, it has to be directed on the bigger picture resulting from all the legal sources related to the theme taken into account, a complex set of information that is often difficult to identify, retrieve and gather in the same context. ” [113]

In fact, the natural connections between legal data (which are partly imposed by precedent and partly through consideration and citation of legislation, professional sources and other materials) means that the domain of legal information does lend itself to network-based analysis and visualisation. Letteri et al propose a system called *Knowlex* which seeks to tie together different sources of law and associated data through visualisation [113]. The system allows a user to specify a legislative measure from statute as a root search term. The web-based application then provides two forms of visual analysis which seek to demonstrate the connectivity between different sources of legal information that are relevant to the initial search query. This is presented in two ways: as an interactive node graph depicting the properties of relevant documents; and as a zoomable treemap which attempts to codify and display the topics concerned in a particular query together with the evolution of legal literature on the point of interest over time.

Another experiment along the same lines is the *Lexmex* project [184]. This is an online system developed in France to map relationships between provisions of and information associated with the French Civil Code. *Lexmex* displays data

sources as network diagrams with nodes and edges. Different texts are connected in the graph if they mention, modify or create one another. The sizes of nodes vary depending on the number of connections each has with other nodes in the graph. The colours used in the diagram correspond to groups of authorities which concern specific paragraphs of the Civil Code, with different colours pointing to different specific paragraphs. The platform allows for interactive navigation through the network.

One of the goals of any legal information system which seeks to visualise information about precedent and the content of case reports should be to bring the user close to the important content of the case report itself. *Lexmex* is a good if limited demonstration of this idea. Too many existing proposals for visualisation posit a particular view of data as the ultimate step in an activity. It is suggested that visualisation should be used to make sense of unstructured, complex information but that it should facilitate and maintain a search pathway to the full text of reports. In this way, the user themselves is the ultimate arbiter of significance and importance. In Ravel Law, there is an artificial separation between text and visualisation which means that reading case law is divorced from the heatmap view. An important legal skill which must be engendered in students is to take texts and to quickly ascertain the critical elements of them so that they may be dealt with. This ability is not taught or developed through differently-siloed, abstracted and isolated views of the text.

3.8 Collaboration environments for lawyers

“Collaboration is a common term in many industries, but a relatively new concept in the legal sector...It promotes innovation, creates capacity, manages risk, and drives quality and efficiency. So, if the business case is so compelling, why is collaboration in the legal industry not more widespread? The truth is, because it doesn’t form part of our DNA. We’re trained as adversaries and brought up in a siloed and competitive culture which recognises and rewards individual contribution.” [152]

The processes of legal research and preparing cases for court are highly collaborative ventures. Research and legal drafting involves teams of associates, senior lawyers, junior lawyers, trainee lawyers, paralegals and partners. These

3. RELATED WORK

stakeholders work together to discover relevant information and to create various written documents and other collateral which will ultimately form the basis for submissions that are delivered by counsel in court. The final presentation of a case in front of the judge also involves multiple counsel who work together and deliver different parts of an argument and submission.

The introduction of computer hardware and computer software into this environment places the case preparation endeavour firmly in the domain of Computer-Supported Collaborative Work (CSCW). Bowers and Benford's general definition can be used for orientation here - "*[i]n its most general form, CSCW examines the possibilities and effects of technological support for humans involved in collaborative group communication and work processes*" [20].

The traditional matrix which is central to the domain of computer-supported collaborative work separates collaborative working environments into different categories. They can involve synchronous and collocated collaboration; synchronous but remote collaboration; asynchronous collocated collaboration and asynchronous remote collaboration. Thus the challenge of facilitating collaboration with computers concerns a range of effort to produce tools that bring people together, to create "*shared workspaces*" [88], to integrate suites of existing tools and to generally promote shared awareness in teams during the course of work activities.

In the legal domain, the awareness question becomes partly an issue of enabling senior lawyers or teachers to know that junior colleagues or students are thinking and working in plausible directions. There are often many different ways to apply the same cases and items of legislation to the facts of a scenario. The problem is non-trivial because case transcripts and statutes are not structured or demarcated according to their intended purpose or final outcome. Post-facto abstracts, headnotes and markup traditionally require input from legally-qualified writers who analyse and systematise the case through their knowledge and experience. The goal of awareness tracking and supervision in collaborative work has been described as aspiring to a "*What I Understand Is What You Understand*" model for work activity [158].

"As lawyers, we did not return to Skype or instant messaging as a writing tool. Our writing practises seemed to lend themselves more to asynchronous (not occurring at the same time) rather than real-time collaboration. However,

it's also important to note that we never returned to our previous practice of one of us writing the first draft in Microsoft Word and sending it to the other as an email attachment. In large part, the reason was that the Writely collaborative online wordprocessing tool arrived on the Internet.” [92]

In their 2018 book, *The Lawyer's Guide To Collaboration Tools and Technologies*, Kennedy and Mighell chart the development of general purpose tools which enable people to work together [92]. They then consider how these platforms can be applied by lawyers in their working practises. A set of guidelines for the successful use of group-based technology in the legal environment is produced. The authors advocate the use of tools ranging from track changes in stand-alone versions of Microsoft Word; Microsoft Office 365 and Google Docs for collaborative document drafting; Skype and Slack for instant messaging in groups; Microsoft Sharepoint for content publishing and sharing on intranets; Google Calendar and other products for scheduling; and the various content publishing and markup facilities available in modern versions of Adobe Acrobat. Once again, a jigsaw of different tools with varying interfaces and limited interoperability is the outcome.

An examination over some time of activity and posts in technology forums for lawyers, such as *The Legal IT Information Network* on the LinkedIn business platform, indicates that attention on collaborative working technology is heavily orientated towards practice management, the mechanics of maintaining case files with inputs from multiple lawyers and the consequent financial activities of billing clients and managing lawyer-client relationships. There is relatively little discussion in either academic or professional circles about the creation of specialist products for lawyers and law students which enable and promote collaborative working especially for legal research.

Reed Elsevier do offer a version of their *LexisLibrary* legal information software which integrates with Microsoft Office 365. This enables legal research and drafting activities to take place in a single, integrated environment and within the interface of the word processor or any other component of the Office platform, including PowerPoint. Thomson offer a similar plug-in called *Drafting Assistant* to integrate some elements of *Westlaw* into Microsoft Office. These approaches have the side-effect of tying the law practice in to expensive licensing arrangements with both Lexis (or Westlaw) and Microsoft.

There has been much more interest in domain-specific platforms and tools for

group working in other professions. Eighteen years ago, for example, [154] presented the results and conclusions from a trial of a shared patient record system in coordinating heterogeneous work amongst groups of doctors with some level of shared purpose in a collocated environment. He discusses the concept, creation and application of environments to allow for common understanding of information amongst groups of physicians, nurses and pharmacists who are engaged with the same patient but who all have different goals and immediate priorities. The later work of Heath et al [75] also contributes to the understanding of collaboration technologies in the medical field. The analogy of medical information to legal materials is superficially appropriate, in that patient data is ultimately parsed and constructed into diagnostic outcomes and treatment pathways. The rest of this thesis examines the need for bespoke software development in the legal domain which enables and promotes collaboration in an integrated environment for document drafting and legal research.

3.9 Conclusion

Computer-based products for legal information retrieval have been available for decades. They are now ubiquitous and are replacing printed law libraries at an accelerating rate. Changes to the way in which case reports and legislation were published at the start of this century facilitated reliable and comprehensive access to online repositories of legal information. The electronic legal information market is dominated by a duopoly between two large publishers, Thomson Reuters and Reed Elsevier. This stifles innovation although content licensing agreements have led to some diversification and the creation of alternative products in the English legal domain. **Contribution C1.2** of this thesis (from the outline in Chapter 1) provides a novel tactical basis for addressing the lack of diversity in the legal research software ecosystem by providing a tightly-integrated open source platform that can be freely forked for different applications and which is based upon open access legal data.

The move to online repositories of legal information has prioritised information coverage and completeness over usability and effective curation. This thesis proposes that future developments in the sector should focus on user requirements and evaluations of the benefits and problems with existing tools. The most detailed and broad studies about how lawyers work pre-date the introduction of electronic legal resources, or certainly their widespread use. This thesis seeks to provide

a modern contribution in the same area which examines how lawyers and law students work and specifically what contribution and barriers computers and electronic information-seeking present. The results of this work will then be applied particularly to the topics of legal information discovery and legal research. **Contribution C1.2** of this thesis provides a legal information platform which seeks to strike a balance between content integration and effective software integration with a focus on usability.

Many commentators agree that the exponential growth in different sources of legal information (particularly in an online setting) is eroding legal research skills. This skills gap is increasingly being noticed and criticised by the courts who warn strongly against the inadequate preparation of legal cases. One goal of this thesis is to ascertain whether a research skills gap exists under English law and to explore the role that technology plays in ameliorating or exacerbating it. **Contribution C2.1** and **Contribution C2.2** seek to provide a starting point for fostering higher levels of research skill in early-stage lawyers by prioritising the development of linguistic analysis competencies in the lawyer as they work with legal sources.

It is suggested that the application of techniques and design paradigms from the domain of corpus linguistics can lead to more effective information retrieval products for lawyers. Although digital corpora have been available for decades, their creation from and application to legal texts is a relatively recent development. Much of the activity in applying corpus linguistics to the law centres on proposals to use empirical linguists as expert witnesses who can testify about the ordinary meaning of language. It is suggested that corpora of case law reports and statutes can also be used to elucidate legal treatment under the doctrine of precedent. The aim is to train lawyers effectively in linguistic analysis rather than to reduce or replace their roles in the legal process. **Contribution C2.1** and **Contribution C2.2** represent the first time that corpus linguistics, corpora and associated interrogation interfaces have been applied to a tool designed for use by lawyers themselves.

Most corpus managers implement some form of the KeyWord In Context paradigm to present search results. This is effective for a trained linguist but alternatives like DoubleTree may be easier for laypeople to understand. They also promote knowledge about linguistic variation around search query nodes.

The move towards different visualisations for corpus search results could sit well with the broader effort to make law more visual and easier to understand.

The connectivity of sources through precedent and statutory interpretation by the courts is a characteristic of legal data which lends it to presentation and exploration in network and tree-based diagrams that can be interactive. **Contribution C3.1** is an attempt to automatically present information about the linkages between legal cases and the nature and significance of these links based upon precedent in an approachable visual layout. **Contribution C3.2** proposes an additional tree-based visualisation that can augment existing techniques for looking at language and particularly at linguistic variation which dictates partly how the nuances of the law are expressed.

It is clear that lawyers must be able to collaborate with others and to work in groups effectively, both in training and in practice. However, the skills required to achieve this proficiency may not be sufficiently emphasised in traditional training programmes at university and elsewhere. Various general purpose tools, from Skype to Microsoft Sharepoint, can help to create shared workspaces and common understanding online. These tools have been applied to the legal sector but more information about their efficacy and uptake in the sector is needed. There also appears to be relatively little effort towards producing specialist collaboration tools for lawyers. These findings form the summary of **Contribution C1.1** of this thesis.

II

PART II

SETTING THE SCENE



CHAPTER FOUR

LEGAL RESEARCH

4.1 Thesis process

This chapter contributes to answering the main research question in this thesis by addressing the following steps of the process outlined in Table 1.2:

- **P1 - Consider how lawyers work in conducting legal research for case preparation.**
- **P2 - Consider the role that technology plays in facilitating or hindering effective working practice.**

These steps will be conducted through analysing the results of three separate studies that are triangulated together in order to provide a comprehensive account of the process of legal research using computers. The objective is to understand problems which are introduced by technology into the legal research process which can be subsequently addressed by the outputs of this thesis.

4.2 Introduction

The first step in the research presented in this thesis is to better understand how lawyers work, what role technology plays in their activities and how computers facilitate and hinder effective working practice. A particular focus is taken on legal research, case preparation and simulated educational contexts. These are all areas where skills deficits have been identified in the United States. Part of the

motivation here was to identify whether a similar issue of low or inappropriate skill levels could be seen in an English law context.

The results presented in this chapter are triangulated from three separate studies of legal working practice. Firstly, a contextual inquiry was conducted with students who were preparing moot cases for presentation in a simulated court environment. Education was the starting point because previous work shows that collaboration techniques and computer-supported tool choice are formed when students are training to become practising lawyers [5].

It was then important to understand how these initial findings from the student group related to professional practice. A set of interviews was conducted with solicitors who were also involved in judging moot cases. Some of the solicitors involved were solicitor-advocates who also presented cases on a regular basis in the courts. This meant that the participants had knowledge of current legal practice in both the profession and in education.

Finally, it was desirable to understand how generalisable and accurate the findings in education and practice were across different areas of the legal profession. An online survey was designed and distributed about collaboration techniques and the role that computers play in that to a larger cohort of practising lawyers with different specialisms. The responses to this survey allow conclusions to be drawn about the broad nature of collaboration, group working and computer-based tool choice in different facets of the legal domain.

4.3 Methodology

4.3.1 Contextual inquiry

The contextual inquiry was conducted over two days between Monday 26th and Tuesday 27th January 2015 at a School of Law in a major British university. The inquiry consisted of two interview and observation sessions, one on each day, which lasted for seventy minutes each. See Appendix A, Section A.9 for the interview questions. Both sessions were recorded using audio and video capture. The first session consisted of an interview with the participants. This was semi-structured and based on a script of questions which was distributed in advance. In the second session, a master-apprentice relationship was established. The interviewer watched the students as they worked and asked questions about

their activities from time to time. The students used the session as a normal part of their case preparation time for two different mooted competitions. The goal of the master-apprentice relationship was to identify tasks in the preparation activities and barriers to effective working practice without breaking the flow of the work or interfering with standard preparation practice. In conducting the contextual inquiry, we used the framework for contextual inquiry and analysis proposed by Hartson and Pyla in [71]. See Section 4.4 of this chapter for more details about the data collection and analysis methodologies that were used.

There were seven participants on the first day for the initial interview. This dropped to four participants on the second day for the observation session. All of the participants on the second day had been present for the initial interview session. The four participants in the observation session split naturally into two separate mooted teams. One team was composed of two second year students studying a joint honours degree in English and Scottish law. The other team was composed of two third year students who were reading single honours English law. Each team was addressing a different legal brief which had been provided to them as a word-processed document by the moot organisers in advance of the sessions.

4.3.2 Solicitor interviews

Interviews were conducted with three senior practising lawyers separately over the course of a day on 26th September 2016. All the participants in this exercise worked for the same major UK law firm. Each of the interviews lasted for an hour, giving a total of three hours of data to analyse and to draw conclusions from. The interviews took place in a meeting room at the offices of the law firm where the participants worked. The exchanges were recorded with audio capturing equipment. A pre-questionnaire document had been constructed in advance by the researcher and the initial period of each interview involved going through the questions on this form in order to gather background information about the participant. This covered their legal experience, time qualified and how often they were typically involved in preparing for litigation. Each interview was then conducted face-to-face between the researcher and a single lawyer at a time. See Appendix A, Section A.10 for the interview script.

The three participants in the interviews were all qualified solicitors. Each of them had a traditional legal background with an undergraduate degree in law from

a UK university as a first qualification. They had all then trained further in law school before undertaking a period of employment as trainee solicitors in various UK law firms. [IP1] had been a qualified solicitor for ten years, was male and was thirty-six years of age. [IP2] had been a qualified solicitor for six years, was male and was thirty-three years of age. [IP3] had been a qualified solicitor for twenty years, was female and was forty-three years of age. None of the respondents had any formal training in information technology.

4.3.3 Lawyer survey

A survey was distributed to legal mailing lists, professional LinkedIn groups, legal officers, practising lawyers and a journal for legal practitioners. The questions were published online. An anonymous web link to the survey was distributed amongst the different sources of potential respondents. Screening questions were included to allow for a focus on respondents who were involved regularly in legal case preparation and legal research. The survey consisted of thirteen questions, some of which demanded multiple-choice answers and some of which asked for free-form textual responses. See Appendix A, Section A.11 for the survey script. The purpose of the questions was to elicit information about how the participants worked in teams; how often they worked with others; which computer-based and physical tools they used to facilitate this; and how their working time was split between different types of activity.

The forty survey respondents were mainly working in legal firms but results were also collected from three legal academics. Twenty of the survey participants were general solicitors, eight were partners in a legal practice, two were legal associates, four were paralegals or members of support staff and three were trainee solicitors.

4.4 Analysis procedure

4.4.1 Contextual inquiry

The contextual inquiry interview and observation sessions were recorded both with audio and video equipment so that data could be extracted and used to model the working practice that was evidenced. The first step was to work out what roles each partner in a mooting team assumed whilst they prepared their moot cases for court. The full work roles diagram can be seen in Section A.1. The physical

arrangement of the individual members of the two teams was also recorded and converted into a diagram. This yields important information about how people work together and how physical barriers to collaboration can be created and overcome in the environment. This physical diagram is shown in Section A.2.

The next step was to work out which high-level tasks the participants were engaging in during their moot preparations. This information was extracted both from questions in the initial interview session and through an analysis and tagging of the video record. The high-level tasks were then broken down into lower-level activities and this forms the basis of the hierarchical task analysis which can be seen in Section A.3. The initial interview questions can be seen in Section A.9.

Flow models were created from a tagged version of the mooting session video recording. The resulting diagrams can be seen in Section A.4. The idea here is to capture the communication and coordination relationships between people which exist to accomplish work. The models show how work is divided into formal and informal roles and responsibilities. The flow models are further annotated with barriers that were observed to effective working practice. These barriers are represented on the diagrams as red lightning bolt symbols with descriptive textual labels. This analysis was augmented by the creation of a social model which shows the expectations of different members of a mooting team, the pressures inherent in different work roles and the constraints which are created as a result of social relationships in the groups. It was also possible to produce a consolidated artefact model which demonstrates what the deliverables from each identified work role are and how these artefacts are communicated during the work process. See Appendix A, Sections A.6 and A.5 for the social and artefact models respectively.

The next step involved three separate analyses of the audio transcript of the mooting sessions by different members of the research supervision team. These provided insights into the scope of the mooting preparation problem and allowed for the construction of individual affinity diagrams. These affinity diagrams group topics of relevance and interest that came up during the sessions into three separate wall-sized maps of associated PostIt notes. Each of the analyses were conducted in isolation without communication with other members of the supervision team. The wall diagrams were photographed and then reproduced in the digital images that can be seen in Section A.7. These affinity diagrams were then consolidated into a single model from which duplicate issues were removed.

4.4.2 Solicitor interviews

The solicitor interviews occurred after the contextual inquiry session and so it was possible to focus and refine interview content to get more detail, corroboration and contrasts from the existing information. The questions which formed the basis of the solicitor interviews can be seen in Section A.10. The solicitor interviews were recorded using audio capture equipment. These recordings were then transcribed into text. The audio file was also annotated with fundamental topic information. This data produced a taxonomy of the topics that were covered. This taxonomy can be seen in Table 4.4.

4.4.3 Lawyer survey

The lawyer survey occurred after both the contextual inquiry and the solicitor interviews. Hence, the questions could be refined and focused again. The full transcript of survey questions can be seen in Section A.11. This chapter includes various statistical and quantitative analyses of the data that was collected from the lawyer survey.

4.5 Results

Study	Scope of data	Source	Total resource size
Mooting: work observation	70 minutes per participant	Video & Audio capture	280 minutes
Lawyers: interviews	60 minutes per participant	Audio capture	180 minutes
Lawyers: online survey	40 participants	Online free text and multiple choice answers	520 question responses

Table 4.1: The separate studies which provide data that is analysed in this chapter, together with the scope of each study.

The sources of data which form the basis of this chapter are described in Table 4.1. Results are organised first by topic (tasks, activities, collaboration and tools) and, within those categories, by the study (survey, interview, mooting observation). Under *Tasks and activities*, findings are collected which relate to the composition

and characteristics of critical legal research tasks. *Collaboration* refers to findings about how multiple people exchange information or how they coordinate with each other. *Tools* refers to issues such as the choice of hardware and software for completing work, its integration and the suitability of available facilities.

4.5.1 Tasks and activities

This section characterises the general areas of legal research and preparing for litigation. The results provide a description of the tasks and activities involved in legal research for case preparation from the three main sources of data (the survey, the interviews and the mooting observation). For additional details on the outputs of the contextual inquiry, please refer to Appendix A.

4.5.1.1 Survey

The results from the survey of lawyers show that preparing for litigation, which includes preparing materials that others will use for litigation, makes up a large part of the working time of people in the legal profession. Overall, the respondents spent 47.5% of their working time preparing for litigation, with some of them spending up to 90% of their time on this general task. The amount of time that legal professionals spend preparing cases for court depends on their seniority and job title, with more senior people (e.g. partners) spending less time than more junior professionals (such as paralegals). See Table 4.2.

Job Role	Average time spent in preparing for litigation
Legal academic	20%
Trainee solicitor	20%
Paralegal/support staff	70%
General solicitor	46%
Associate	80%
Legal partner	43%

Table 4.2: Percentage time spent on preparing for litigation by job role.

The survey respondents were asked to detail the top five types of activity that are part of preparing for litigation. Their answers, categorised into nine classes, showed strong variation between individuals. Communication and project management, case research, negotiation and evaluation, and preparing court documents are cited most often as common tasks (the full list is in Table 4.3).

Types of activities in descending order of participants' mentions
1. Communication and Project Management
2. Case research
3. Negotiation and evaluation of other party's position
4. Preparing court documents
5. Witness identification and recruitment
6. Document discovery
7. Giving legal advice
8. Establishing the counter-case
9. Attending court

Table 4.3: Sub-activities of preparing for litigation, in order of frequency, from the survey data.

4.5.1.2 Interviews

The solicitors in the interviews confirm the data from the survey about the importance of the general activity of preparing for litigation. They also offer some detail about the most important sub-tasks within preparing for litigation. For example, [IP1] shares that *"I'd say at the outset maybe 70% of my time is spent on factual investigation, 30% legal interpretation"*. [IP2] describes how *"finding the state of the law on something is 10% of my job but legal research is important and is a big part of the nuts and bolts of working on a case"*. [IP3] explains that *"I probably spend half my time analysing the law and half my time analysing factual material"*.

The interviews provide additional information about the various work roles that different types of legal professional take on (see Figure 4.1). These roles are closely related to the sub-tasks described in Table 4.3, with most professionals having to take up multiple roles. However, certain tasks are strongly associated with specific titles in the profession. For example, solicitors work with barristers and advocates to search for relevant cases that establish precedent but solicitors are mainly responsible for managing the project, whereas barristers or advocates are mainly in charge of presenting cases in court. Assistants (e.g. paralegals) might carry out more mundane or repetitive tasks in all of these areas. There is also a practical split in the work roles between legal trainees, junior lawyers and senior lawyers, associates and partners within the firm. The full list of roles and responsibilities that we found through the analysis of the interviews is detailed in Appendix A, Section A.12. Due to the distribution of roles, most cases involve intense collaboration by two or more professionals with the client, often from within the same legal firm. [IP2] said that *"I work very rarely just on my own because*

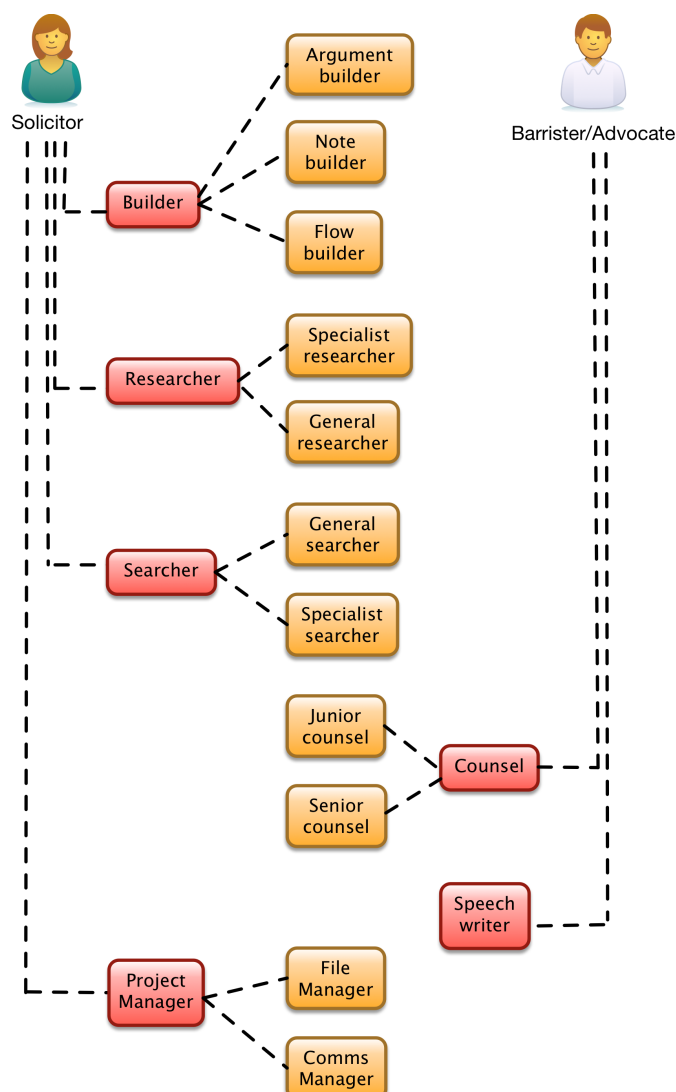


Figure 4.1: Work roles diagram from the interviews with solicitors. Discontinuous lines indicate typical connections between the type of role and the professional title. Assistants that are not solicitors or barristers/advocates can do any of the roles and are not shown in the figure (for clarity).

any case will always involve the client. So you have to collaborate with the client and their witnesses, the people who work there, who know about the case. You are immediately part of a team." Although in some types of case (e.g. personal injury claims) the majority of the work is undertaken by a single junior lawyer, all of the respondents said that they work with other people in most legal matters very frequently.

To map the work that lawyers carry out, responses from the interviews were used to detail recurrent tasks, sub-tasks, time consuming work and issues that are part

4. LEGAL RESEARCH

of legal practice. These are detailed in Table 4.4. Notice that this table is different to and independent from Table 4.3, since the results come from different data sources, yet they are closely related in terms of dominant topics and time commitment.

High-level tasks	Low-level tasks
Case Preparation	Document Discovery; Factual analysis; Finding cases; Time management; Transforming notes into strategy
Collaboration	Communicating with stakeholders; Managing multiple inputs to the case file; Revising and discussing documents; Maintaining confidentiality; Waiting for information; People management; Strategy creation and management; Selecting team members according to budget
Working with the Court	Developing court document submissions; Managing information formats; Answering judicial requirements; Creating and maintaining a transcript
Managing multiple digital devices	Moving documents between devices; Formatting submissions for different form factors; Maintaining data security; Auditing the flow of information
Catering for different levels of experience	Agree on IT infrastructure; Delegating tasks based on legal qualifications; Delegating tasks based upon specialist legal knowledge; Creating a team with relevant experience for a case

Finding legal Information	Managing search results from different tools; Executing agreed search queries
Using specialist legal software	Coordinating and maintaining the case file; Keeping track of case activities; Updating search tasks based on communications from the team; Adapting previous case skeletons to current case
Preparing for court appearances	Interfacing with barristers and advocates; Developing a courtroom strategy; Aligning client interest with court strategy; Adjusting strategy as the case proceeds; Reacting to the opposition performance; Discussing and negotiating settlement; Managing barrister fees within budget
Taking notes	Making sense of handwritten notes; Creating digital documents from handwritten notes; Maintaining notes for audit and policy review; Creating file notes from working documents; Extracting legal advice from notes

Managing formal documents	Creating suitable formal documents based on strategy; Creating suitable formal documents based on budget; Bringing disparate contributions together; Ensuring that the case file is accurate and comprehensive
---------------------------	--

Table 4.4: Low-level taxonomy of tasks and topics in the legal profession from interviews

4.5.1.3 Mooting

A hierarchical task analysis was conducted from audio and video records of the mooting observation sessions. The availability of video data enabled an analysis of the activities of the participants in more detail. Ten high-level tasks were extracted that the mooting groups carried out to be able to undertake their cases (Table 4.5). These include things like reading the case brief and looking for relevant legislation.

Separately, the types of low-level activity which the mooting students spent their time on were also elucidated (Table 4.6). These low-level activities were subdivided into broad classes when appropriate. For example, the composition of documents will normally include splitting and merging of information (Table 4.6). The two sets of work effort (tasks and activities) are related through a further analysis that lists which activities are necessary to carry out the tasks (the “decomposition” column in Table 4.5).

In a further step, an analysis of the generic tasks involved in producing work products was completed in order to determine which accounted for the most time during the mooting preparation exercise. This timing information was derived from an annotation of the video record of the sessions as detailed in Table 4.7.

From these results, it is clear that the most time consuming activities have to do with writing by hand (**Externalisation – Writing**). This includes noting the key elements of data found online and documenting discussions and thinking. These are both endeavours which often support subsequent knowledge synthesis. Verbal brainstorming within the team was a key method of understanding relevant cases and other data and then working out how to apply this information in the novel

No	High-level task	Decomposition
1	Read the case brief	1a, 2a, 2b, 2e, 4a, 4b, 1b
2	Identify seed cases	1a, 4a, 6a, 5
3	Look for relevant legislation	1a, 4a, 2d, 2e, 6a, 5, 7b, 7a
4	Split the case brief	1a, 7a, 2f, 2a, 2e
5	Search for related cases	1a, 6a, 5, 4a
6	Build an argument	6a, 5, 7a, 7b, 4a, 2a, 2b, 2f, 2e
7	Identify the counter-argument	6a, 5, 7a, 1a, 2f, 4a, 2a, 2d, 2e, 7b, 2c
8	Identify relevant journal articles	6a, 5, 7b, 2f
9	Prepare a speech	1a, 3a, 6a, 7a, 7b, 2a, 2b, 2c, 2f, 4a, 3b, 4c, 5, 4b, 2e, 6a
10	Prepare a rebuttal	1a, 7a, 4a, 2f, 3b, 4c, 4b, 6a

Table 4.5: High level task descriptions from mooting observation.

Low-level activities	Activity type
1. Read	a) Text
	b) Graphics/visuals
2. Knowledge formation and Synthesis	a) Discuss
	b) Argue/reasoning
	c) Review/summarise
	d) Brainstorm
	e) Decide
	f) Selection and filtering
3. Composition	a) Splitting
	b) Merging
4. Externalisation	a) Writing
	b) Diagramming
	c) Typing
5. Transfer/store	
6. Interact with software	a) Interact
7. Search	a) Close reading (within documents)
	b) Distant reading (between documents)

Table 4.6: Low level activities and sub-activities derived from mooting observation

circumstances of the mooting problem. Digital externalisation using software tools was comparatively rare. Indeed, most of the activity on a computer was concerned with finding information which would then be physically parsed and filtered by hand.

Flow diagrams were constructed from the contextual inquiry data. These show how information is created, parsed and filtered during the preparation of a

4. LEGAL RESEARCH

	Activity	Duration (seconds)
	Composition - Merging	785
	Composition - Splitting	640
	Externalisation - Diagramming	373
	Externalisation - Typing	421
	Externalisation - Writing	3609
	Interact with software - Interact	581
	Knowledge Formation & Synthesis - Argue and reasoning	599
	Knowledge Formation & Synthesis - Brainstorm	2041
	Knowledge Formation & Synthesis - Decide	309
	Knowledge Formation & Synthesis - Discuss	420
	Knowledge Formation & Synthesis - Review and summarise	261
	Knowledge Formation & Synthesis - Selection and filtering	248
	Read - Graphics or Visuals	65
	Read - Text	1738
	Search - Close reading within documents	1968
	Search - Distant reading between documents	1175
	Transfer or Store	1121

Table 4.7: Task timings from mooting observation. The colours represent encodings of activities in Figure 4.6

moot legal case and how it moves between key actors. The diagrams bring an appreciation of the role that legal information plays in the case preparation process and how filtering of data results in new work products. The ten separate flow diagrams (one for each high-level task in Table 4.5) are available in Appendix A, Section A.4. From these, a consolidated diagram of transitions between the different types of activities is shown in Figure 4.2.

The diagram in Figure 4.2 shows that a dominant switch pattern exists from externalisation (writing notes and drawing diagrams) to knowledge formation and synthesis and vice versa. Participants often used their notes to synthesise new ideas. The primary digital activities that were repetitive here were searching for information (using either generic or specialist tools) and transferring that data

for reading on screen. An exclusive relationship can be seen between interacting with software and the transfer or store state. This means that the role of digital tools was restricted to search and retrieval tasks and downloading information for future reading.

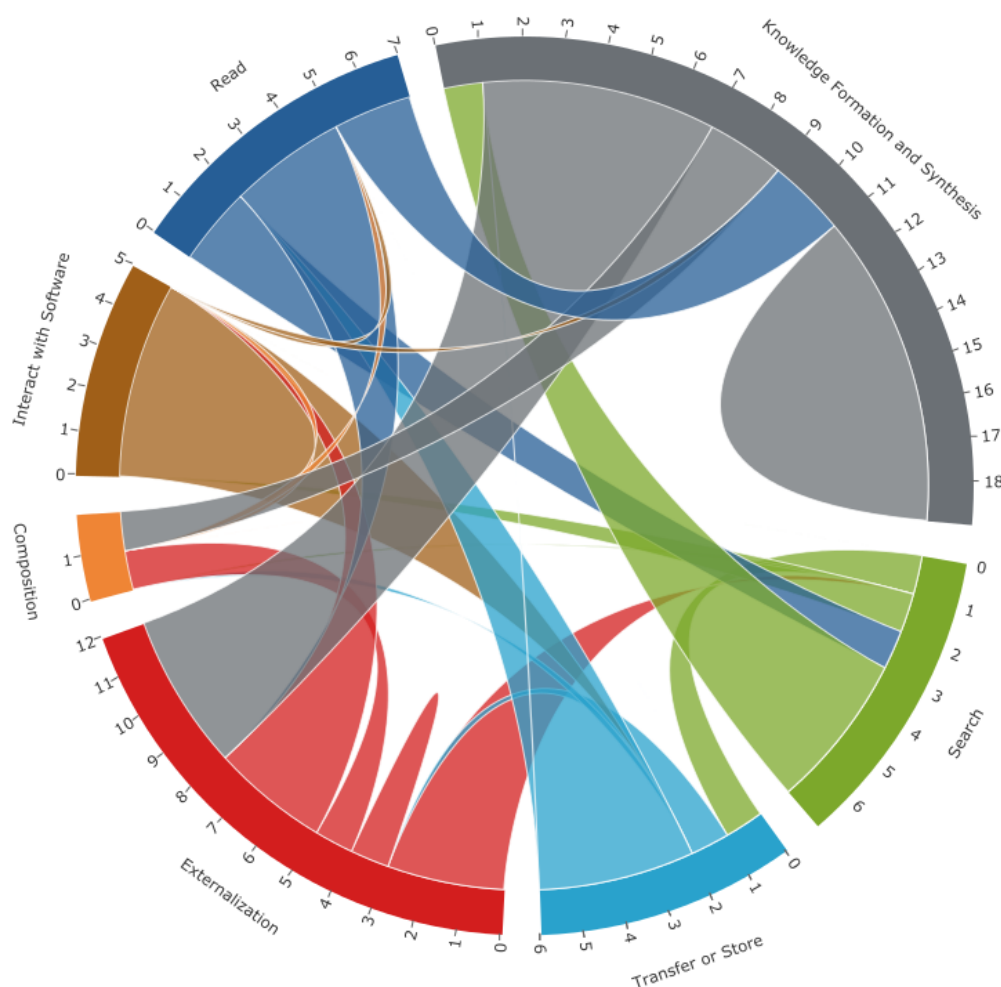


Figure 4.2: The activity transition chord diagram. The nodes around the outside of the circle are the low-level generic tasks from Table 4.7. The size of each node shows the total number of transitions to that state that we recorded. The edges are the number of transitions from a source state to a destination state. The colour of each edge shows the source state that the transitions come from. Loops within the same state are shown as edges that originate and terminate in the same state.

From the flow diagrams and the transition model in Figure 4.2, five high-level work roles were identified that were assumed by the two members of each mootng team at different times during their work. These high-level roles were split into several more specific designations, which are associated with the generation of particular documents (Figure 4.3). The descriptions of the sub-roles are detailed

in Table 4.8. Notice that, although most roles emerge from the activities required by the job, two of them (*junior counsel* and *senior counsel*) are mandated by legal procedure.

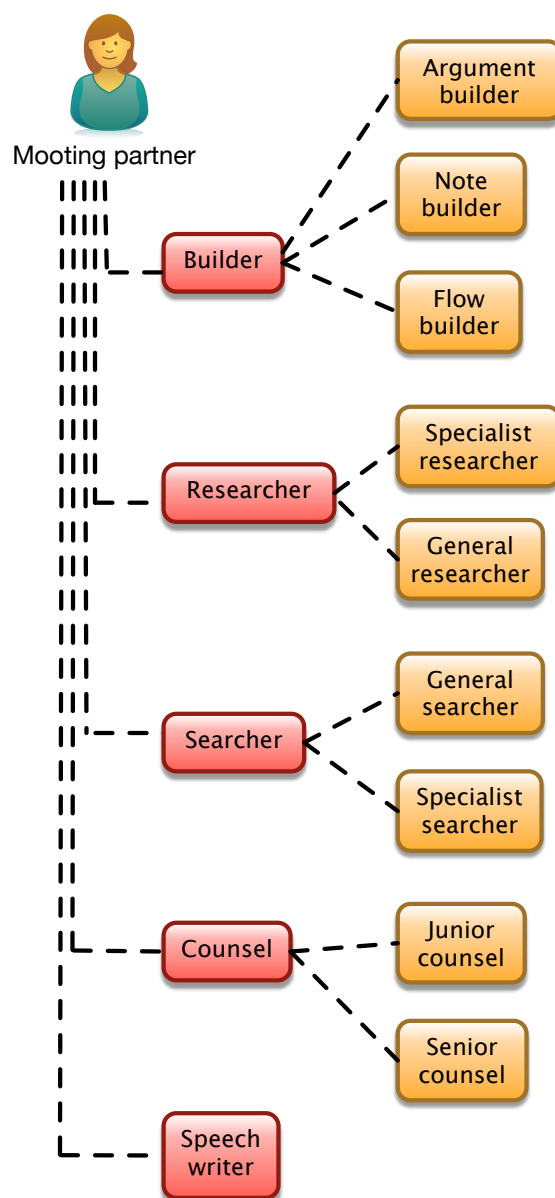


Figure 4.3: The work roles model from mooting observation

Roles and responsibilities
Builder
High-level classification for roles which involve work product output.
Argument builder

Writes argument skeletons and final legal brief to be put forward in court.
Note builder
Takes notes of important information during the preparation process. Brings disparate notes together into coherent information.
Flow builder
Prepares fact and transaction diagrams to make sense of the problem question.
Researcher
High-level classification for roles which involve researching an issue in law.
General researcher
Consults stored output from general purpose tools like Wikipedia and Google to create a treatment of important information.
Specialist researcher
Uses stored output from specialist tools like HeinOnline and Westlaw to create a treatment of important information.
Searcher
High-level classification for roles which involve searching for and storing important information.
General searcher
Uses general purpose information tools like Wikipedia and Google to find and then store case reports, legislation and other legal documents for future research.
Specialist searcher
Uses specialist information tools like Lexis Library and Westlaw to find and then store case reports, legislation and other legal documents for future research.
Counsel
High-level classification for the only mandated split of responsibility between the members of a mootng team.
Senior counsel
Responsible for either starting the submission to the court or finishing it, depending upon jurisdiction. Can be responsible also for the rebuttal if applicable.
Junior counsel

Responsible for either starting the submission to the court or finishing it, depending upon jurisdiction. Can be responsible also for the rebuttal if applicable.
Speech writer
High level classification with the responsibility of bringing information and arguments together into a coherent speech for delivery in front of the judge.

Table 4.8: Work roles and sub-roles identified in the mootng observation

4.5.1.4 Summary and Cross-methodology Results: Tasks and Activities

Activity in the process of preparing legal cases is often collaborative. Lawyers and other actors have to work together with others - be they clients, the court, witnesses and domain experts or the opposition - during multiple phases of case preparation. Individual lawyers and support personnel take on many different roles and work tasks are generally divided into granular low-level activities in order to create deliverables which contribute to producing final submissions to the court. Some of the low-level activities that take a large amount of time are intrinsic to the task (reading text, synthesising new knowledge). However, significant time is also spent on activities like note-taking which appears to often be sub-optimal.

This is because taking notes usually includes direct copying of information like case citations and quotes from case reports and legislation. Time is also spent on activities which are overheads caused by technology, like interacting with software to search for information and then storing case reports and legislation for later synthesis. Low-level activities are characterised by frequent context switching between looking for relevant content, taking notes and generating new content.

4.5.2 Collaboration

The results detailed in the previous section indicate that collaboration is common and critical in litigation. It was desirable to explore these findings more thoroughly from all three data sources. It is important to understand what collaboration means in a legal context at a detailed level because working patterns are complex. They are informed both by tradition and mandated procedure and they have been impacted by the introduction of information technology.

4.5.2.1 Survey

The survey responses provide data about the nature of collaboration in multiple firms and in different specialisms. Lawyers must work with others in the same firm and with outside specialists like domain experts, witnesses and barristers in bringing a case to court. Initially the survey group was asked how much of their working time was spent collaborating with others.

The mean percentage of time spent collaborating was 77%, minimum 36%, maximum 100% across the 40 respondents. The survey data shows that the time spent collaborating in individual responses varies substantially. To unpack what is understood by the term “collaboration”, respondents were asked to specify the top five activities in their work which need collaboration. The full survey data on this question (25 responses, 100 sentences) can be found in Appendix A, Section A.13. These responses allowed for the creation of a high-level taxonomy of collaborative activities, presented in Table 4.9. That table shows the collaborative activities that members of the lawyer cohort are often engaged in, ranked by frequency of selection.

High level collaboration tasks in descending order of frequency of selection	
1. Internal general communication (email/telephone/meeting)	100%
2. External general communication (email/telephone/meeting)	100%
3. Strategic instruction and discussions	47.5%
4. Document production	37.5%
5. Witness and evidence preparation	35%
6. Procedural discussions	30%
7. Case research	22.5%
8. Settlement and fee negotiation	10%
9. Attending court	7.5%

Table 4.9: High level task ontology of collaboration tasks.

All of the respondents mentioned work which fits into tasks 1 and 2 (internal and external communication with email and other forms of distributed group working). It can be seen from these results that email drives collaboration among the survey respondents. The scope of legal work is shared between experienced and inexperienced staff, senior and junior staff and by legally qualified and unqualified staff. Table 4.10 shows how much time is spent collaborating based on legal job type. The results demonstrate that junior staff collaborate more. They also do more legal research. In Section 4.5.2.2, the interviews are considered which

4. LEGAL RESEARCH

highlight the gravitation of legal research work towards unqualified and junior members of the legal team.

Job role	Amount of time spent collaborating
Legal academic	0%
Trainee solicitor	71%
Paralegal/support staff	80.5%
General solicitor	61%
Associate	70%
Legal partner	60%

Table 4.10: Percentage of working time spent collaborating by legal job description

Between the most senior and junior workers, the results show that associates collaborate more than legal partners but less than the most junior staff. It is important to note that the questions related to case preparation as a work activity, hence the 3 legal academic responses give a 0% figure for time spent collaborating. This reflects how these respondents are involved in the law.

4.5.2.2 Interviews

Collaboration in the professional legal domain is largely dictated by budget and whether a particular case is defined as “*big ticket business*” versus “*volume litigation*” [IP 2]. Volume cases are generally undertaken by a single lawyer (although, as was evident in Section 4.5.1.2, this also requires collaboration with a range of other people such as clients and the opposition. Larger cases will require a team of people which necessitates more coordination and collaboration. [IP 2] said that “*the three most important things...in big ticket litigation are...collaboration so that all the right people are involved at the right time, it’s document management, and it’s probably also project management, making sure all the different parts are operating well at the same time.*”. [IP3] said that “*...I always find the delegation piece quite a difficult one to do. If you’re working integrally with someone else on a case, it’s fine for you both to know every detail of the case.*”. [IP 1] stated that “*we tend to have trainees...doing the sort of mundane stuff. The more strategy-driven aspects of the case would be something to be determined between senior lawyers and perhaps advocates and counsel.*”. These results relate not only to collaboration but a range of other issues that are covered in the Discussion (Section 4.6).

	same time (synchronous)	different time (asynchronous)
same place (colocated)	Mooting preparation Brainstorming	Mooting preparation Document creation Sharing case links
different place (remote)	Legal Practice Billing Client file updates	Legal Practice Case preparation by email

Figure 4.4: The CSCW Matrix for legal case preparation.

With reference to Figure 4.4, respondents in the interviews highlighted that collaboration was often distributed and not colocated. [IP3] said that email drove collaboration “*hugely - 99 percent. And it is also probably the most problematic element of working...the correspondence traffic is all driven by email and the volume of that is enormous in every case*”. [IP1] said that coordination between the group of lawyers on a case would “*happen electronically, by email [and] when we meet face to face, someone will have their email on a screen for reference or we’d have a printout of the relevant emails*”. [IP2] stated that he would “*take a laptop with me when working with witnesses because you often need to refer to important emails that have been sent when deciding...what to cover with them*”.

4.5.2.3 Mooting

The spatial arrangement of collaborators in co-located situations has been of interest to Human Computer Interaction researchers for a long time - for example, in [49]. It is relevant to the design of support tools and environments [102]. The two mooting groups under investigation arranged themselves as shown in Figure 4.5.

The physical arrangement from Figure 4.5 is not specific to legal collaboration, but

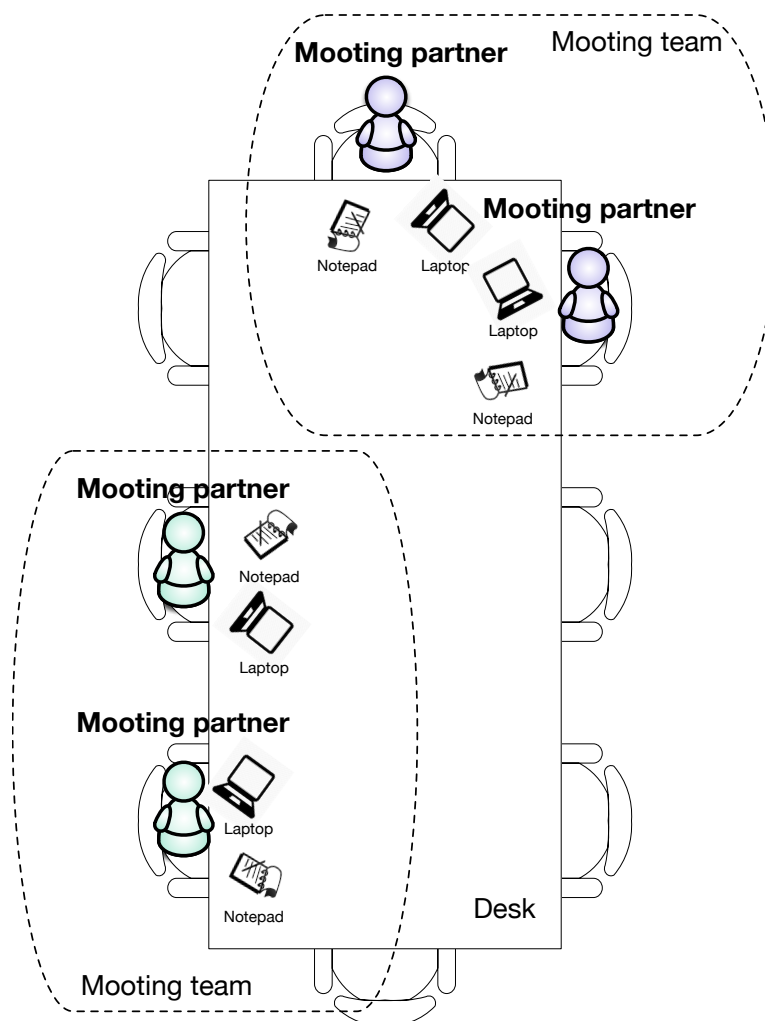


Figure 4.5: The physical arrangement of mooters across the two sessions.

exhibits patterns that are important to understanding how co-located collaboration in this domain is affected by technology. The positioning that can be seen in Figure 4.5 is conditioned by the fact that people in the legal profession need to use a combination of digital tools (usually a laptop or a tablet) and other tools (mostly pen and paper). Working together on a laptop is difficult, since it is not easy to share a laptop screen or a tablet if the collaborators are sitting in separate, relatively remote chairs. In fact, the results show that the laptop screens themselves become barriers because the backs of the laptops serve to cut workers off from one another.

It should be noted that there were infrequent changes of position in the mootings group which itself resulted in relatively few instances of real-time collaboration.

In fact, synchronous collaboration was mainly affected through textual communication. There was some use of Facebook Messenger to share materials. This tended to replace more direct forms of synchronous collaboration and resource sharing. There was some glancing at the screen of an individual's mootting partner. However, the seating arrangement made this difficult and unnatural. Glancing was usually facilitated by moving a laptop so that the other team member could see it, before moving the laptop back to its original position. The collaboration diagram in Figure 4.6 shows where the participants in each mootting team talked to each other in order to facilitate collaboration, and also where one member of a team glanced at the screen of their partner.

It was found that most instances of verbal communication between team members were exchanges that were not work-related. Discussion that was to do with the mootting activity was mainly conducted through instant messaging. This separation occurred because the work was formalised and needed to be recorded in some way. However, the majority of work effort was individual and siloed. It was not discussed in real-time within the teams. This reflects the fact that a lot of notes and other output, particularly at early stages of the case preparation process, are "half-baked" or tentative ideas which the partners do not feel confident to share until they are more fully expressed and tested through legal research.

Another barrier to real-time collaboration arose because the different legal information tools had varying levels of coverage. A number of instances were recorded where one partner in a team found a case which was relevant but the other member of the team could not find it. This variability of search outcome happened where different platforms were being used within the team and also where both partners were using the same software.

The video recording of the activity data was analysed to describe how collaboration took place. Figure 4.6 shows the pattern of activity types and indicates when closely-coupled synchronous collaboration on the same task took place. Most of the work products were created individually rather than jointly. An example here occurred early in the observation exercise when one team debated whether a duty to perform under a contract in their brief was a *fiduciary duty* or not. They decided that it was not but the information and resulting analysis was collated by each member separately.

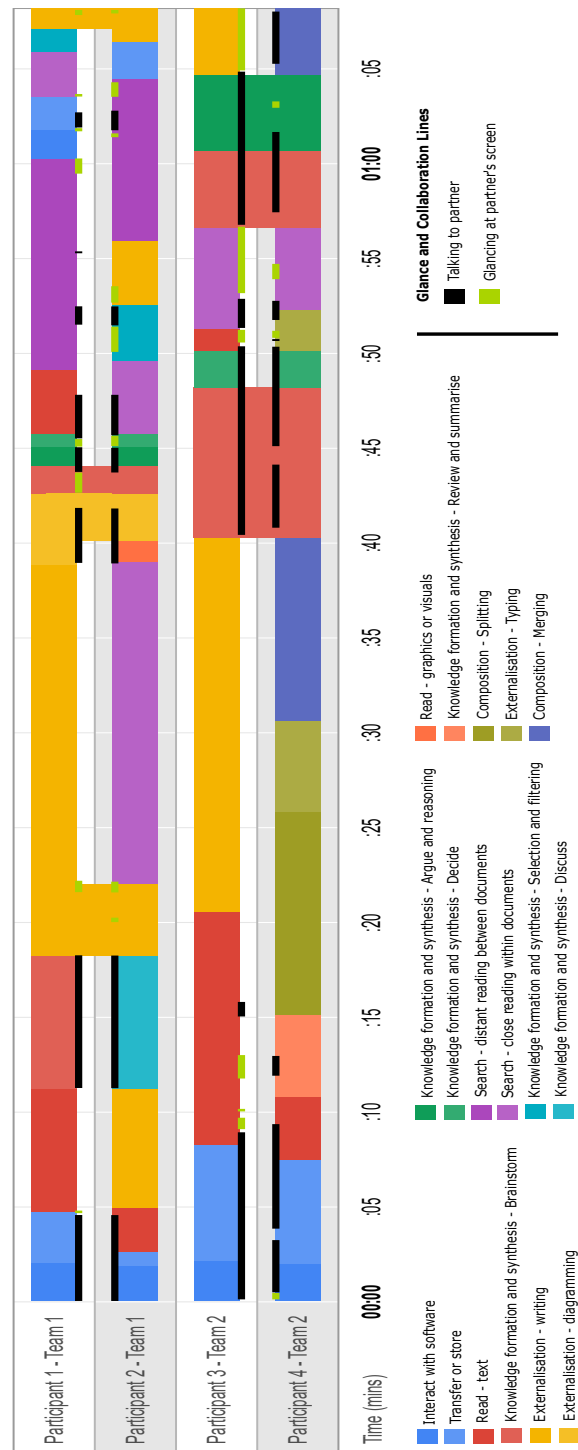


Figure 4.6: The collaboration diagram. There is an activity bar for each person which shows which of the low-level activities the participant was engaged in. The key for the colours is given in the legend and corresponds to the low-level tasks described in Table 4.7. A black line along the bottom of the activity bar for a participant shows where they talked to their partner. A green bar indicates when they glanced at their partner's laptop screen. Activity bars cross where real-time collaboration took place.

4.5.2.4 Summary and cross-methodology results: Collaboration

Collaboration is pervasive throughout the legal research process and is usually asynchronous and remote, orchestrated through email. When collaboration is collocated, it tends to be asynchronous. This is true both in legal practice and in legal education, where the participants in moot teams gather in the same room but their work is largely independent. The products of this effort are then brought together later. Students are more likely to make use of technology for instant messaging than qualified lawyers are. Where collaboration is collocated and synchronous, the use of common materials is rare. The reasons for this are unclear but technology certainly does not facilitate collocated and synchronous collaboration at present.

4.5.3 Tools

4.5.3.1 Survey

The survey respondents used a mix of electronic equipment, digital tools (e.g. software and systems) and physical tools (e.g. pen and paper, diagramming) in case preparation and legal research. The results on tool use, shown in Table 4.11, are based on free form textual answers to question 8 of the survey.

Tool	Number of responses
Electronic mail	18
Specialist legal information tools (Westlaw, PLC)	11
Telephone calls	9
Physical letters	7
Specialist case file management software	6
Physical meetings	5
General word-processing software (Microsoft Word)	3
Virtual meetings (video conferencing)	2
Electronic calendar	1

Table 4.11: Ranked order of tool use by frequency of response (derived from survey)

In terms of the digital resources employed, Table 4.11 indicates that this is dominated by email and specialist legal information tools (e.g. Westlaw). However, traditional tools which support verbal and physical interactions (e.g. telephone and letters) remain widespread. A surprising result here is the lack of integrated

solutions in use, like *Lexis for Office 365*. There is no evidence that these solutions have penetrated into the profession in the results.

The preference among the survey respondents is to use a number of different, simple and non-integrated tools during the course of legal research. Specialist legal information tools are used in isolation to research and to find relevant legal precedents. The drafting process starts when data from these tools is summarised and synthesised using pen-and-paper notes. Email maintains a dominant position in the collection and distribution of legal information within teams of lawyers. The most integrated part of the preparation process, as evidenced in Table 4.11, involves specialist case management and file maintenance software. These tools centralise records of emails sent and received. The goal here is to streamline and simplify the processes of billing and auditing case progress.

The act of externalising information is well understood in forming and synthesising what one currently understands only as tacit knowledge [139]. As such, the survey cohort was asked to explain how they split their time and work activities between using digital tools and pen and paper. There was a heavy preference for taking pen-and-paper notes over using digital tools. 75% of the survey respondents tended to use pen and paper in preference to computing devices and software. The full results here are shown in Table 4.12.

Response	Percentage of responses
I exclusively use pen and paper to take notes	16.67%
I mostly use pen and paper, but I might take notes using digital devices - e.g. a laptop and word processor	58.33%
I mostly use digital devices - e.g. a laptop and word processor - but occasionally I use pen and paper	20.83%
I only use digital devices - e.g. a laptop and word processor	4.17%
I don't take notes	0%
Total	100%

Table 4.12: Notetaking split between physical and digital (from survey)

The results in Table 4.12 lead to the next question: how pervasive is note-taking in bringing a legal case to court? It was found that the average amount of working time spent taking notes in the survey cohort was 50%, up to a maximum of 90%. To drill down into these results, the individual responses to this question were

analysed in order to find out if there is a correlation between increased levels of physical note-taking and those jobs which typically involve legal research. The results here are shown in Table 4.13.

Job Title	Average % of time spent notetaking
Legal academic	10%
Trainee solicitor	75%
Paralegal/support staff	30%
General solicitor	59%
Associate	72%
Legal partner	41%

Table 4.13: Average percentage of time spent taking notes by job title (from survey)

The results show that junior qualified lawyers spend more time on average taking notes than others in the profession. Those tasked with legal research take more notes than those who typically deal with strategy. Notes are often taken partly using pen-and-paper and partly on computing devices. This means that notes relating to the same case exist in multiple media which can lead to fragmentation of important information.

The survey data also shows that legal professionals depend on general purpose computer-based tools for day-to-day work. This is especially true for activities such as document creation and administration. For the purposes of legal research, specialist electronic legal information tools are pervasive. The use of these tools is limited to activity 6 (case research) in Table 4.3. Video conferencing is a relatively recent departure but it is becoming commonplace. 90% of the respondents selected general tools like email as their most commonly used software. The most dominant specialist tool was *Westlaw* for legal research.

Collaboration is driven by...	Percentage
Brainstorming	24.53%
Strategy outcomes from meetings	47.17%
Skype or video conferencing	20.75%
Dropbox	3.77%
Microsoft SharePoint	3.77%
Google Docs	0%
Total	100%

Table 4.14: Tools that drive collaboration in the profession.

There is evidence of a lack of penetration for cloud-based collaboration tools in our survey responses. This is evidenced in Table 4.14 by the low level of selections that we recorded for tools like Dropbox (less than 4% uptake), Google Docs (0% uptake) and Microsoft Sharepoint (less than 4% uptake). The respondents also had the option to specify other tools in this question and there were no cloud platforms in that data either.

Finally, the survey cohort was asked how satisfied they were with the range of tools (both digital and physical) which are available to facilitate collaboration. There was a high degree of variability in this response but the average satisfaction rating was 64%. The lowest satisfaction score was 6%. In order to find out which levels of the profession were happiest with existing collaboration capabilities, the results are broken down further by job title. Legal academics and non-qualified staff are least satisfied with existing collaboration facilities. The highest levels of satisfaction were among general solicitors and trainee solicitors. The full results here are presented in Table 4.15.

Job Title	Average % satisfaction
Legal academic	50%
Trainee solicitor	70%
Paralegal/support staff	53%
General solicitor	70%
Associate	70%
Legal partner	77%

Table 4.15: Satisfaction levels by job description (level of seniority). Standard deviation in these results was 24.2%.

4.5.3.2 Interviews

[IP1] described himself as a “*tech geek*” who felt that he had a good general facility with computers. [IP2] stated that he was experienced with information technology “*to the extent that I know what’s out there and can use different tools well in my work*” whilst [IP3] stated that she had “*a good understanding of technology for someone of my age.*”

All of the solicitors that were interviewed used a variety of computing devices in their work. They each had a desktop PC in the office which was secured to their desk. This computer was the platform that they used most of all in the course of their work, for all the activities that are described in this chapter.

Two of the interview respondents made some use of a laptop and editing software in order to take notes. This choice of digital resources was driven mainly by financial considerations in that clients would not welcome charges for transcription of handwritten materials. [IP2] said *"it is important to use reliable software to take notes and to write reports...simple applications which were proven to be stable, because everyone knows that Word can crash and has had the experience of losing a document which hasn't been saved"*. [IP 3] tended *"to take all my notes and to make key reports by hand before typing up important materials"*.

[IP 3] stated that note-taking was an important technique for planning legal cases and for auditing activities on the behalf of a client at a later date. Drafting documents at all stages of the legal process also started with sketching and note-taking. *"I start with putting together an executive summary and then the meat of the report and get a structure of subheadings for particular aspects of the case or aspects of the claim"*. [IP2] said that the number of notes that they would take depended largely on how prepared and organised an individual client was. *"Some clients are less switched on and you really need to draw the information out of them, so that involves lots of conference calls or client meetings. Working final documents up from notes that can be lodged with the court is important. That takes up a lot of time and involves a lot of re-writing and refinement."*

[IP 3] said that she would have lots of *"notes, just working notes, that I've scribbled. Some of my working notes are such a mess that you wouldn't believe. They'll just be totally incomprehensible to anyone else but I can look at them and see what I was thinking at the time. And I will retain them so that people can see them. But a lot of our clients are really looking for the advice being quite quick and very nimble and flexible and at a low cost."* In that scenario, she would work on a note which comprised a set of bullet points which could be transferred into an email quickly and easily.

Regarding their use of generalist or specialist tools, [IP 2] said that the most significant recent addition to the software tools in the firm was Microsoft Lync. All physical telephones had been removed and calls were conducted over Lync. This participant described the introduction of Lync as *"transformative"* because it allowed working with external parties more effectively - *"I can show them expert drawings, schematics, take them through the nub of the case and the issues that are in dispute"*. [IP 3] stated that working practice and collaboration within the firm was still *"driven hugely by email"* and cloud tools were not used *"because I don't think that our IT infrastructure could handle it and there would be serious security concerns"*

around client confidentiality.” [IP 1] stated that they needed to “find a way of reliably archiving Conversation History entries from Lync as part of the client file, as currently there is no method for adding these records to the case record.” This is an indication that lawyers need to archive and centralise a number of multi-modal sources (for later retrieval) within the same case bundle, generally organised by client.

When asked about their satisfaction with existing computer tools, [IP 3] said that she felt that computerisation was *“inevitable and there is no point trying to live in the past”*. However, she stated that there was now a *“Google generation of trainees and junior lawyers who would research topics just by plugging ad-hoc queries into a search engine”*. She said that *“...newer lawyers find that - they don’t really worry so much about [researching] precedents and the authority of decisions. They don’t really understand that as well as people would have done in the past. Because they’re so used, I think, to having a kind of technological view of searching as opposed to a case precedent basis for searching.”* [IP 2] said that *“...not only is the quality of research diminishing. I mark assessments and over the nine-year period that I have been doing it...the quality of work that you get from the students is going down every year. The skills of legal research really are diminishing. Students seem to arrive as trainees in law firms with very poor legal research skills. You ask them to research a point of law...and they’re coming back with a sort of bastardised summary of what was online.”*

Both [IP2] and [IP3] were involved in judging mooting competitions and in marking student assignments at local universities. This means that they had an insight into the quality and nature of work which students had been submitting for assessment over a period of time. [IP3] and [IP2] both highlighted the problem of information overload as the greatest challenge that they faced when using computer-based legal information platforms. [IP3] said that *“information overload is a huge issue in legal work. It is made worse by a lack of search history records. I always open multiple cases in tabs in my browser and then I lose them again when I accidentally click the close button. There needs to be better ways of knowing what information you have seen and which of these sources were important during a research session.”*

4.5.3.3 Mooting

Each student brought their personal laptop to the inquiry sessions. All but one of the participants had a mobile phone. The university provided access to three specialist digital information tools for lawyers: *Lexis Library*, *Westlaw* and *Solcara*. The students all used their laptops primarily during the observation sessions. Use

of Facebook Messenger was evidenced for communication. One student used a legislation search and retrieval app called *iLegal* to find statutes on their tablet. The use of mobile devices was limited, however, in comparison to the amount of time spent on laptops.

The students preferred taking physical notes to using digital applications. [MP1T1] said “...notepads don’t run out of batteries”. In fact, only one member of one team used a word processor during the inquiry. The students said that word processors were too complicated compared to taking physical notes.

The students said that activities like diagramming on a computer necessitated a lot of effort to achieve even simple layouts which were trivial to create by hand. Visual layouts were seen as a good way to make sense of information which is nuanced and precise. One student said that drawing a diagram with pen and paper was “the only good way to see how money and goods flow between parties” [MP2T1]. The use of paper diagrams extended to taking “the content of moot answers and distilling them for revision purposes. Moot questions often cover topics which may come up partly in our exams, just like tutorial questions” [MP3T2]. This participant stated that he used diagrams of case nodes around a central topic node to contain information about key cases and legislation which related to a legal area. However, [MP4T2] stated that they did not use diagrams for revision because “I am not a visual learner.”

Extensive use of general information tools including Wikipedia, Google search, Google Scholar and an open-access legal resources website for students called *eLawResources* was noted amongst this group. The students said that general-purpose information tools were frowned on by the moot court. However, they made use of them to gain an initial understanding of the key topics in a case brief before turning to specialised legal information repositories for detailed work on the problem. [MP1T1] said that “you would never use Wikipedia or quote from it in a moot but it is very good as a general tool. You can look up a case on Wikipedia and get a good general description of what the case is about, although I have seen summaries on there that are wrong or clearly haven’t been written by a lawyer”.

The split between general and specialist tools favoured specialist legal information tools for research purposes with general tools used for information communication both within and between the two groups. [MP4T2] said “I use Facebook Messenger to keep in touch with [MP3T2] and to share important case links. I don’t use the built in chat at all. I’m familiar with Messenger and...I like the way it works, also because there is

a message history that you...can look back at."

There were significant usability concerns expressed by the students about specialist legal information tools. They noted that the tools were all designed to facilitate searching for individual cases or statutes. *"They are much less effective if you want to search on an area of law, like you can with Wikipedia"* [MP2T1]. They said that they found it difficult to obtain relevant information on topic searches like *strict liability*. Another issue arose from the fact that case citations are not all covered equally. There is a hierarchy of identification based upon the court that has delivered a judgement and the number of times that a case has subsequently been referred to in other cases. *"It's confusing, especially in first year, because you are never sure whether to use a citation, or which citation to use. Is there a difference between using an All England Law Reports citation or a neutral citation?"* [MP4T2].

The students also said that ranking and filtration techniques in digital legal tools leave the user with lots to do. *"The search facility is not as exact as it could be. There is too much information given in response to a search - the computer chucks information at you"* [MP3T2]. For example, *"there is no automatic segregation...between cases and legislation in the search...so you get a mixture of cases and statutes...when you might only be looking for one [type of source]"* [MP2T1]. Also, *"...the full report for a case...can be hundreds of pages long...when you are only interested in the head notes and the judgement...but you have to navigate through the whole thing"* [MP1T1].

Another issue that the participants highlighted, which is connected to the point above, is that search refinement in Westlaw is predominantly post-query. *"It would be much better if you could say what you're looking for first - getting a large set of results is daunting...there is a psychological barrier when you first get a list of hundreds of results"* [MP1T1]. This issue is exacerbated by the fact that search facilities within documents are limited online. All of the students preferred to download case reports that seemed relevant in PDF format so that they could search within the text using the PDF viewer. [MP4T2] said that *"searching through the online text...is taxing...because it can be badly formatted and difficult to read...and the PDF is much easier to deal with"*. However, *"...because [the PDF viewer] just searches for the words that you type in, you have to look at the context...yourself to find out if the paragraph or whatever is really relevant...which takes a lot of time"* [MP2T1]. [MP1T1] said that *"the issue of having to switch [from the core search environment] to search in a document or to type is really why I prefer taking notes by hand."*

It is noted that the participants found the issue of context switching frustrating. They adopted a workflow where they would search for multiple cases in sequence, downloading the PDF reports of each in turn, before starting a PDF viewer to search within the text of the group of reports that they had previously stored. This strategy minimised the number of concurrent switches between different applications. Another area of obvious difficulty was finding cases reliably. There were several instances where one member of a team was unable to find a particular case whereas their partner found it. Sometimes this search problem happened on the same platform and sometimes one partner found a case on Lexis that their colleague could not find on Westlaw. [MP1T1] said that *"it is often particularly difficult to find...older cases reliably. They are usually there...but they may be hard to locate, for whatever reason."*

Finally, the participants highlighted a lack of flexibility in the user interfaces of specialist tools. This meant that it was difficult to read large amounts of information on screen. One participant said that *"when you are up at two o'clock in the morning working, [Westlaw] is almost impossible to work with"* [MP4T2]. Another participant said that she *"liked the way that Westlaw looked"* so the problem here is that there are a lack of options for adapting the user interface to suit individual preferences.

4.5.3.4 Summary and Cross-methodology Results: Tools

The findings show that note-taking is an important activity in legal case preparation. Notes are usually made by hand using pen and paper. This is because digital tools for writing involve another context switch away from search and retrieval of information in work roles that already require frequent changes of attention, as shown in Figure 4.2. The preference is for taking hand-written notes before synthesising them and then converting important content to typed documents later. These typed documents are often emails. The software tools chosen for transcribing notes tend to be simple because they offer the highest levels of reliability and the lowest additional learning curve.

Lawyers use email messages and tools like Windows Notepad to transcribe their notes rather than specialist solutions like *Lexis for Microsoft Word* or even general-purpose word processors. Email is a dominant tool for orchestrating collaborative work and also for creating important notes which relate directly to the progress of a case. This is partly because emails provide a specific audit trail for activity

on a case. However, email does not integrate well with other physical or software tools outside of case file archiving. Security and confidentiality considerations are paramount in the legal sector and they are a strong inhibitor on the uptake and adoption of new tools. Newly-qualified lawyers and novice support workers transfer their general styles of searching for information on the Internet into legal tools. This behaviour is perceived by senior practitioners to create a skills deficit in the work of their junior colleagues.

4.6 Discussion

The survey of 40 legal professionals and, to a lesser extent, the interviews with practising lawyers confirm a number of key points about preparing for litigation and conducting legal research. First, this is a core activity in the legal sector. It is time-consuming and involves people of varying experience and seniority in multiple roles. These facts reinforce the need for further research in this area. However, such research faces many barriers. Legal professionals are often busy, stressed and overworked [161], and they are very aware of the economic impact of their time [80]).

The characterisation of the tasks and activities that legal professionals carry out at work also shows a job that requires significant amounts of *communication* and *coordination* with others, even for small legal cases. This places case preparation as a type of work that, despite having significant time pressures, does not have the real-time monitoring demands of other sectors (e.g. [9] and [74]). All the sources also point to a complex combination of activities where coordination and communication with others, information search and the synthesis of new ideas and documents are heavily interleaved.

Legal professionals are generally not formally educated in the use of information technology. The specialist software that they use is devoted almost exclusively to information search and retrieval instead of covering and integrating a wider range of their work activities. *“Lawyers are not always open or motivated to train themselves on the use of legal technology applications or even basic applications such as MS Word or MS Excel. While mobile apps...have simple interfaces and features designed to be intuitive, legal technology applications and productivity software does not. [A lawyer’s] understanding and use of technologies is often simplistic and fails to leverage tools and functionality for real-time savings as part of their legal practice”* [65].

4.6.1 Modes and Tools of Collaboration and Communication

Some of the more specific findings in the results section have to do with the way in which collaboration takes place. The interviews and survey revealed that most of the interchanges of information involved in preparing for litigation are carried out through e-mail. Therefore, most collaboration takes place in a distributed and asynchronous way (quadrant 4 of the CSCW matrix in Figure 4.4).

The nature of information in legal collaboration (e.g. precise citations and references, potentially contradictory pieces of evidence, possible lines of argumentation) do not seem particularly well supported in the tools that most legal professionals use. Quoting or copying citations, references to cases and links is cumbersome through e-mail. It requires the added work of copying and pasting between the e-mail client, text editors and legal search tools. There is also a need for significant discipline and effort in synchronising the state of important documents on top of record keeping. Although some tools exist that might help with this (e.g. cloud based tools specifically geared towards the legal professions such as NetDocuments and iManage), it is clear from the interviews that their penetration is still relatively low (see Section 4.5.3).

The data collection methods used in this chapter were not designed to answer why collaboration is still driven by e-mail. However, it can be speculated that the reason is one of or a combination of the following: a) collaboration across firm boundaries makes it difficult to agree on a particular tool; b) legal professionals are not generally trained in information technology and therefore they do not have sufficient understanding of how relatively new technologies might provide value [67]; c) existing tools do not address the collaborative needs of legal professionals better than simple e-mail or are cumbersome and difficult to use; and d) e-mail is perceived as a safe current practice compared to other technologies that place information away from the purview of the firm. The current data along with that collected by others somewhat support a), since professionals have to collaborate with others beyond their own firm (see Section 4.5.3.2). For b), [108] highlights the multi-layered and complex collaboration chain in any commercial legal work. Regarding c), usability of even the well-established search tools is considered sub-optimal (Section 4.5.3.3) and d), alternatives to e-mail are perceived as less secure.

It is somewhat surprising that collaboration is asynchronous, since the moot

observations suggest that there is value in synchronous communication, even when tasks and roles can be easily split among collaborators. For example, the real-time sharing of case links from legal information platforms saved time in the preparation exercise and enabled both partners to continue with certainty. Other authors have found that collaboration in document production is driven asynchronously which, whilst introducing some problems, also brings benefits. Chandler [28] states that asynchronous distribution “*allows team members to continue working together with a reduced level of interpersonal interaction, thereby increasing their productivity*”.

However, it is not possible to directly answer why asynchronous collaboration is dominant in legal work. The reasons might not just be technical, but: a) distributed asynchronous collaboration is perceived as more efficient, as found above; b) distributed asynchronous collaboration places fewer constraints on how legal professionals allocate their time (finding common times in already busy schedules); c) there is no appropriate support in current tools; and d) tools supporting distributed synchronous or co-located synchronous collaboration are not yet well known enough among lawyers.

4.6.2 Note taking

Another key finding is that note taking is extensive and takes place using pen and paper with a much smaller reliance on digital tools. It was found that legal practitioners use notes for three main purposes: as a way to support their own memory of the large amounts of information involved in the legal process, as a facilitator to transfer information between different systems or parts of the process, and to keep personal records of the process in an auditable resource.

This form of externalisation essentially forms the “glue” that enables legal practitioners to carry out their work. Generic unstructured note-taking might be preferred due to its simplicity and flexibility. The activity of note-taking also supports the process of cognition itself, regardless of whether the notes are ever checked again. “[It has been] proposed that note taking functions in either or both of two ways: as a technique that enhances “encoding” of passage material and/or as a means of storing material externally (“external storage” hypothesis). The encoding hypothesis suggests that the act of taking notes results in a transformation of passage material. The precise nature of the transformation has not been fully specified, but it likely involves some processing beyond verbatim learning...The external storage hypothesis, on the other hand,

indicates that notes are taken to store passage information likely in verbatim fashion for later use for recall purposes” [155].

4.6.3 Search and Supervision

Specialised legal search is probably the best digitally supported activity in preparing for litigation. There are multiple tools in the market which facilitate a welcome trend towards publishing of reports online. Legal professionals are thus able to access millions of relevant documents from the courts that, in the past, would have only been accessible through expensive legal libraries that had to be visited in person. Although this results in undeniable savings in time and, according to the interviews, money spent by the legal firms on paper documents, it does introduce a new set of problems of a different type.

Probably the most trivial is that the usability of these tools seems to be deficient. For example, post-query refinement and search facets are inefficient and create a psychological barrier for the user; interfaces are fixed and cannot be organised according to individual preferences; and the maintenance of query histories relies upon simplistic tools like bookmarks in a web browser. It is also clear that the collections that are accessible through different tools are heterogeneous, which forces legal researchers to either use tools with different interfaces or, even worse, to ignore part of the legal literature.

Also problematic, but within the remit of tool design, are issues to do with the way that search is implemented in current legal tools. Most existing software implements structured search by title, date, or case citation. This style of search is alien to many new practitioners who are used to the general internet search style. This might result in deficient searches without the practitioners even noticing that they are missing a significant number of relevant documents. It was observed that mooted students reverted to general Google and Wikipedia search of the Internet which, at the moment, are not complete sources of legal documents. This is corroborated by the interviews, where [IP2] and [IP3] both complained that the legal research skills of newly-qualified trainees are severely lacking. This deficiency impacts the time and efficiency of the overall legal preparation process. New researchers are not systematic in their searches and do not have a good understanding of when they have uncovered sufficient relevant material. Senior practitioners are then forced to closely supervise their searches. This activity in itself is poorly supported by software.

4.6.4 Opportunities and Recommendations

From the results of the three studies, some key recommendations for the future development of software tools to support collaboration and information retrieval in the legal sector can be proposed.

Base case and legislation coverage on open access data: One of the primary inhibitors to product diversification in the legal sector is the difficulty that new entrants have in obtaining legal data upon which to build their tools. Although court transcripts are public documents, lawyers expect them to be formatted as case reports with abstracts, head notes, keywords and other editorial content. This is time consuming and expensive to create. The majority of high-quality legal information remains closed-access and controlled by a small number of established publishers. Tool developers need to start pushing for and using open access data under favourable licensing agreements. They can then invest effort in designing new technology.

An integrated approach to support workflows is more likely to succeed: From Figure 4.2 and from Figure 4.3, it can be seen that there are a large number of transitions required between different low-level activities and between work roles whilst preparing for litigation. Thus novel software tools should concentrate on tight integration in order to support the entirety of the high-level process of preparing cases. This is software integration as distinct from open-ended content integration. The aim is to reduce the number of transitions which necessitate switching between different software tools.

Support synchronous collaboration in the creation of work products: The flow diagrams which were created from the outputs of the contextual inquiry (which are available in Appendix A, Section A.4) demonstrate that most work products in case preparation are text documents. Currently these documents are created individually or through remote, asynchronous collaboration with others. Software creators should attempt to support both asynchronous and synchronous collaboration in teams based around document drafting. Email is currently used as the primary driver of collaboration in the profession but this is really a lowest common denominator. In order to change the current paradigm to better support group working, synchronous and asynchronous collaboration should centre on the document with an integrated platform.

Prioritise note-taking and knowledge synthesis: The transition diagram in

Figure 4.2 shows that lawyers transition between externalisation and knowledge synthesis activities most often during case preparation work. Externalisation tends to mean note-taking or drawing diagrams using pen and paper. As such, a priority in the development of new tools should be the interface between collaboration, note-taking and knowledge synthesis. Reducing the need for transitions between different tools whilst working is a good way to promote more efficient working practice.

Search is critical: Results from the contextual inquiry and the lawyer interviews highlight the problem of information overload as a barrier to effective working practice in the legal sector. This is partly caused because search for legal information is usually implemented through keyword matching in long documents. There are a large number of hits for individual queries where only a small subset of the results are truly relevant. Changing the paradigm for search is therefore a particularly important goal. Many newer legal information platforms prioritise using machine learning and artificial intelligence in order to understand what a user expects in response to particular queries. It is suggested that there should be a separate and parallel focus on providing qualified legal professionals with higher-quality information so that they can make better and quicker decisions about relevance themselves.

Privacy and security are paramount: The results show that, although the adoption of remote cloud-based collaboration and information storage facilities is growing in the legal sector, penetration for these solutions remains low. There is a clear distrust of solutions which abstract storage facilities to remote locations because the security and privacy of client information is essential. Developments like private cloud infrastructure may change this attitude over time but for now it is suggested that any new tools should be designed for internal hosting and direct control in individual law firms.

User interfaces and interaction models must be flexible: The results from the contextual inquiry and the lawyer interviews highlight several usability problems with current legal software tools. These issues centre around fixed and inflexible user interfaces which cannot be customised to individual preference. This is a relatively easy problem to solve but it points to a requirement for a focus on Human Computer Interaction and user experience design for this sector.

4.7 Conclusion

The activity in preparing legal cases is often collaborative. Lawyers and support personnel take on many different work tasks and roles during their preparations. Some of the activities, like externalisation and note-taking, are currently sub-optimal and are conducted in a simplistic manner so that individual productivity is not overtly compromised. Collaboration is usually asynchronous and remote in the profession or asynchronous and co-located in mooting and education. The use of common materials in a synchronous manner is very rare and is limited to glancing intermittently at laptop screens as work proceeds. Technology does not support synchronised and co-located collaboration in the legal sector well at present. Collaboration is instead driven heavily by email.

New entrants into the legal profession are often found to be deficient in their legal research skills by more senior colleagues. This is partly due to the transference of general Internet search techniques into legal information tools. These specialist tools require a more faceted and structured approach to searching. This finding is corroborated by other research but the finding that software tools actively contribute to the skills deficit is a novel result. From the three studies that are reported here and the triangulated set of results, it is recommended that novel software developments for use by lawyers should focus on using open data; on integration to support the whole case preparation workflow; on supporting the collaborative creation of textual documents; on streamlining note-taking and knowledge synthesis using a computer; on implementing new search techniques which prioritise the context of hits; and on flexibility and customisation of user interfaces in tools which are internally hosted.

This chapter has provided **Contribution C1.1** from the outline given in Chapter 1 - a better understanding of how lawyers and law students conduct legal research using software tools. Enabling collaboration between and within groups of lawyers through a usability-focussed set of software integrations forms **Contribution C1.2** and this is addressed in Chapter 5. The use of corpus linguistics, the creation of large-scale living corpora from legal sources and user interfaces which enable lawyers to interrogate a corpus effectively form **Contribution C2.1** and **Contribution C2.2** from the introductory outline. The idea here is that fostering linguistic analysis skills in early-stage legal practitioners can ameliorate problems of information overload and can serve to foster higher levels of competency in legal research activities. These elements of the thesis are described in Chapter 6.

LARC - THE LEGAL RESEARCH AND COLLABORATION PLATFORM

5.1 Thesis process

This chapter contributes to answering the main research question in this thesis by addressing the following steps of the process outlined in Table 1.2:

- **P3 - Consider how open source software can be implemented as an integrated platform for legal research.**
- **P4 - Consider how synchronous and asynchronous collaboration on legal documents can be supported by technology in an auditable manner**

These steps in the process will be addressed through the description of a prototype software platform for legal research, which is called LARC (the **LegAl Research and Collaboration** platform). This software has been designed with reference to the usability and productivity barriers that have been identified with existing platforms in Chapter 4. The effectiveness of the approaches taken will be ascertained by reference to the functionality of an established commercial tool for legal research - see Section 5.13.

5.2 Introduction

In this chapter, the results of the surveys in Chapter 4 will be used to propose and implement a prototype legal research platform. This software is designed to provide an integrated environment for information seeking and document drafting. The idea is to facilitate the creation of work products through models of synchronous and asynchronous collaboration within teams of lawyers. The development and prototyping activities reported here were informed by the consolidated list of serious barriers to effective working practice that were identified with existing legal information platforms - see Table 5.1.

The scope of this work includes developing a back-end infrastructure for the new platform which takes an existing source of open access legal data and then parses it so that it becomes predictable and reliable enough for use in an integrated software application. It also encompasses user interface design and testing based upon a set of initial paper prototypes which can be seen in Appendix B, Section B.1. The interface incorporates a number of detailed visualisations which are included to address specific barriers. These components have been separately designed and prototyped - see Appendix B, Section B.2.

At each stage, design decisions in the prototype software will be justified with reference to the informing usability barriers and the objectives which they imply. Ultimately, a number of key visualisations and other features of the prototype will be evaluated against similar functionality in an existing legal information platform - JustCite by Justis Publishing. That platform is one of the few commercial products which emphasises visualisation of data along the same lines as LARC. The key difference between the two platforms is that JustCite is manually-curated whereas LARC uses automated algorithms to present legal information.

5.3 Current barriers to effective working practice

In Chapter 4, the results of a series of studies were reported which involved legal students and professional lawyers. These were designed to help build an understanding of legal research work from the point where most preferences for research tools are formed. Hartson and Pyla's analysis framework for contextual inquiry and analysis was used here [71]. Next, interviews were conducted with three senior lawyers from a leading law firm. Each of the participants at this stage were involved in judging mooted competitions for local universities. The

5.3. Current barriers to effective working practice

Barriers and description of barriers
1. Pen and paper notes in silos. Extensive use of pen and paper to take notes. Computers very rarely used for note taking. Pen and paper notes are individual and create many sources of isolated information. Notes exist in silos.
2. Context switches inhibit computer use. Lack of integration in software tools. Necessitates switches between different applications. This inhibits productivity and collaboration.
3. Information overload. Search facilities prioritise depth over specificity. Difficult to identify highly relevant material. Many search results difficult to filter post-query. General internet search norms counter-productive.
4. Relevance and hierarchy not shown. Text-heavy search results compromise clarity, appreciation and identification of significance.
5. Difficult to bring notes together. Collaboration is obstructed. Notes on cases need to be integrated into final court documents in isolation from group activity.
6. Synchronous collaboration difficult. Tools do not support working together at the same time. Students tend to be collocated. Professionals tend to be remote. Collaboration currently asynchronous.
7. Email dominant but not ideal. Collaboration driven by email in the profession. Provides simple, instantaneous communication. Adds another information silo. Management is difficult and time consuming.
8. Auditing is difficult. Current tools do not support auditing of research. Senior lawyers and tutors need tools to be able to check and direct research work.
9. General tools are better. Lawyers find general purpose tools easier to use and more powerful, particularly for communication and collaboration. Security and privacy concerns inhibit uptake.

Table 5.1: Key barriers to effective working practice from the contextual inquiry, lawyer interviews and online survey which are reported in Chapter 4.

final step was to design and conduct an online survey which was distributed electronically to a range of lawyers with different specialisms. It was important to know how generalisable the findings from the first two groups were within the profession as a whole. The result is a consolidated list of barriers to effective working practice which are not well addressed by existing tools for legal research - see Table 5.1.

The identification of key barriers enables the derivation of a set of focused design goals and implementation objectives for the LARC platform. See Figure 5.13 for a view of the final prototype interface. The platform is a **document authoring** system for lawyers (*barrier 1*) which facilitates a move away from isolated paper notes. The platform is an **integrated, single context** application which allows for legal research without switching away from the document context (*barrier 2*). **Linguistic analysis and frequency-based search result filtration** is implemented to provide relevant search results from within a large database of legal information (*barrier 3*). The system utilises **information visualisation** to promote an understanding of precedent and the hierarchy of judicial decisions (*barrier 4*). The document authoring functionality is **collaborative** (*barrier 5*), allowing groups of lawyers to work **synchronously in both collocated and distributed environments** (*barrier 6*). LARC includes chat facilities for **messaging between the users** working on a document (*barrier 7*). **Auditing, progress review and direction of effort** by tutors or senior lawyers is central and easy to do (*barrier 8*). The system is based upon **customised and integrated open source software components** so that facilities are standardised and offer an experience as close to general purpose applications as possible (*barrier 9*).

5.4 The overall system architecture

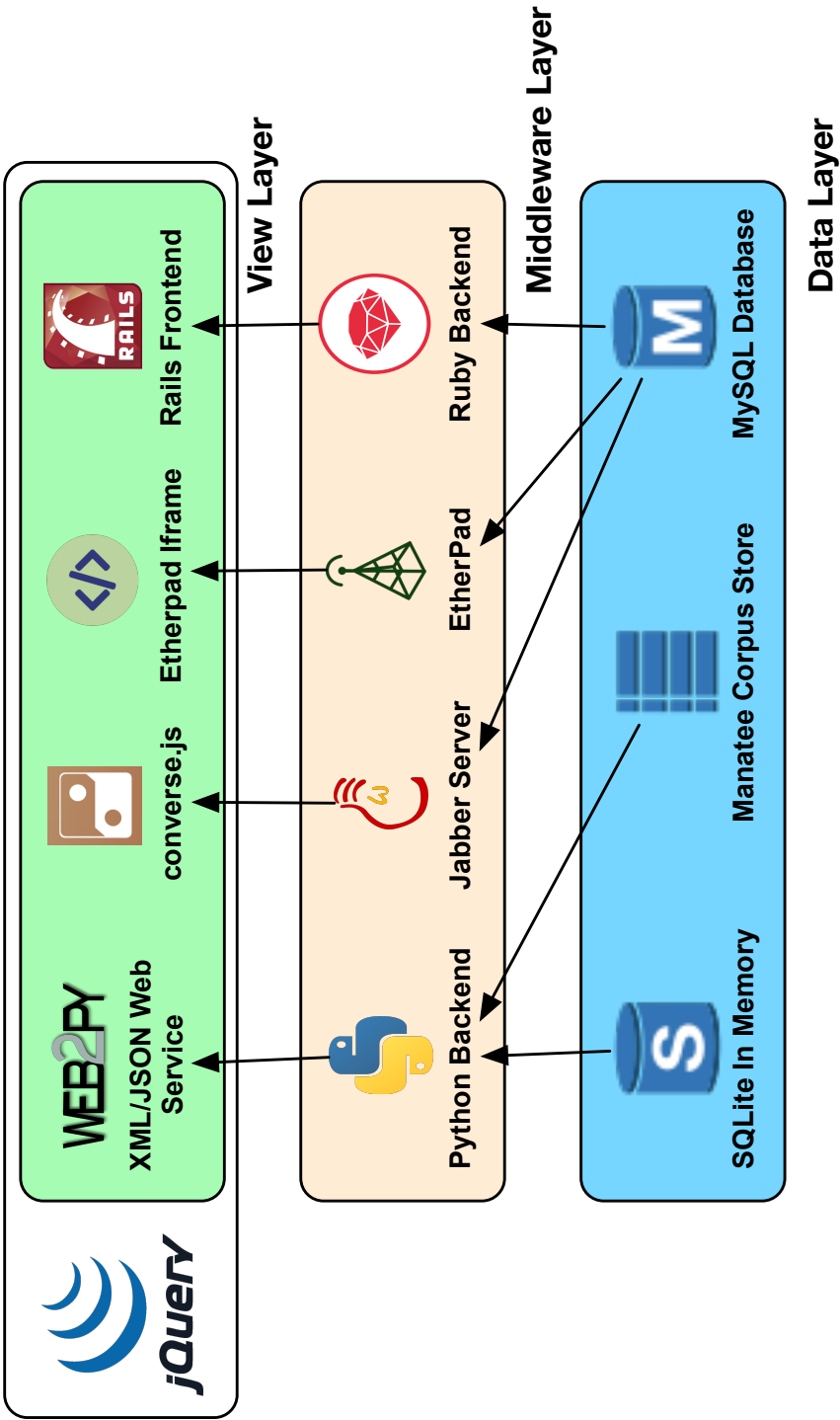


Figure 5.1: The LARC infrastructure diagram.

A fundamental tactical decision was taken to customise and integrate open source software components in LARC. There is also a move towards open access legal data in the United Kingdom at the moment. Nevertheless, the amount and scope of freely-available legal information in this country remains limited in comparison to other jurisdictions. At the same time, small start-up companies which offer new solutions in legal research are being acquired by the large existing monopolies. The future of their contributions in the market is unclear [141]. It is suggested that a platform which is based upon both open access data and open source software can lower the barriers to market entry for 'big data' and legal information experts who wish to provide novel products in this space.

For this research project, a license agreement with the British and Irish Legal Information Institute (BAILII) was secured. This enabled the use of their case law database for research purposes in the LARC project. The collection, parsing, linking and organisation of this data - which contains some 80,000 cases under English law from all levels of the court hierarchy - forms the basis of the LARC infrastructure. Legislation data was provided by The National Archives (TNA) in London. This is XML-formatted content for some 75,000 statutes which had to be similarly parsed and structured for use in LARC. Preparation scripts for transforming HTML content from BAILII and XML from TNA are generally implemented in Ruby.

If linguistic analysis is required at this stage, Python is used instead to facilitate access to libraries including *Scikit* and *NLTK*. A full account of data preparation for the LARC platform is given in Section 5.5. A relational database was chosen to hold the processed information because the relationships and dependency structures that can be implemented here, together with fixed schemas for tables, served to impose an order on information from unstructured sources. See Section 5.6 for more details. The LARC application itself is designed in a Model View Controller pattern using Ruby on Rails. This enabled fast development and prototyping of the system.

The corpus-based language search facilities in the application were implemented using an open source management platform called *NoSketchEngine*. This software works on binary-encoded plain text. There are therefore two separate data sources. The relational database enables attribute-based search and presentation of pre-processed information in the user interface. The disk-based text index holds data for language interrogation. Scripts were prepared to transform the HTML for

each case stored in the database into plain text and then through the preparation process for creating a NoSketchEngine data store. The design and implementation of language search in LARC is described in detail in Chapter 6.

The infrastructure uses well-supported open source toolsets. For example, *Jabber* and *converse.js* are used for in-application chat with data storage to the MySQL database. In fact, any *XMPP* server and front end could be swapped in to the system with minimal effort. A different corpus management solution, such as *CQPWeb*, can be used easily because the preparatory format for most alternatives is the same as that for *NoSketchEngine*. Some of the novelty of LARC rests on the close integration of components that has been achieved to create a bespoke application for legal research.

5.5 Using open access legal data

The data archive published by the British and Irish Legal Information Institute (BAILII) [111] is a large collection of web pages which contain transcripts of important case reports under English law. The site implements a very simple design and layout which is augmented by rudimentary search facilities so that particular cases can be found and read online. Most of the case reports themselves are simple HTML documents and the source code indicates that each document has been exported for web publishing from Microsoft Word.

The legislation archive which is provided by The National Archive is a large collection of XML documents which implements a custom schema for codifying statutory information. The collection has been made available as an open access resource for research and it is updated regularly as new instruments come into force. A significant amount of effort was therefore required in order to build a readable HTML document of each statute from the underlying XML data.

5.5.1 The limited extent of open access legal data under English law

BAILII provides the largest and most coherent source of open access legal data which is currently available under English law. The site publishes most of the more significant judgments which are heard in the courts although it can often take some time for an important report to become available on the site after it

is released. It also provides a good selection of significant historical reports so that the development of precedent over time can be accurately established by consulting the available data set. Discussions with technologists in the legal information sector indicate that some material on the BAILII site comes from court transcripts which are sourced directly from reporters in the courts. This means that the material generally lacks augmented content such as head notes, paragraph references, numbered pages and so on. A license agreement to use the entire BAILII database in the LARC project for research purposes was secured at the start of this research project, without which the work would not have been possible. Amicus Resolution [3], a legal technology company based in Edinburgh, and other companies had previously attempted to license content from BAILII for their own products, but they were not successful in securing an agreement.

“There is some argument whether judges are public servants or not and hence whether their judgments are public sector information or not. In addition, regarding older judgments, the low level of originality required for copyright protection in the UK means that almost all older cases are copyright of either the transcriber or the reporter (or the publisher who commissioned them).”

[82]

Part of the problem here is that there is a vexatious issue of copyright in court reports and the text of legal judgments. As the quote above highlights, it is not clear legally where copyright lies and with which party in the creation and publication process for case reports. The idea that copyright should exist at all in fundamental publications of the legal system which need to be widely disseminated and known in order for the state of the law to be clear under a precedent-based justice system is controversial, especially in modern times when distribution via the Internet is normal and commonplace. Some commentators on the article in [82] point out that a balance between general access and the right to privacy of parties in a legal dispute is not easy to achieve.

However, others maintain a stronger principle that open access should be facilitated by removing the concept of copyright as it applies to case reports and court transcripts. Others contend that copyright arises at source with the Crown who can then make a decision to publish under an open access license agreement, which the state has been willing to do for Supreme Court judgements and legislation in recent years. Nevertheless, this state of affairs means that

legal publishers (including BAILII) tend to more or less jealously guard their sources of legal information and are reluctant to allow for mass re-publication and repurposing of the case reports which they have created.

In fact, many of the case reports on the BAILII website come from material which is prepared and published by the Incorporated Council of Law Reporting for England and Wales (ICLR) [86]. The ICLR was established in 1865 to provide a reputable and complete source of high-quality law reports from case transcripts. The service is relied upon today by lawyers, students and academics as it has attained a status as the only officially recognised purveyor of accurate case reports from the English legal system. The ICLR themselves claim to offer the largest and most comprehensive collection of contemporary and historical legal case reports of the databases available through other information publishers. However, the ICLR archive itself is not open access. Although the organisation is a charity, it charges consumers for access to the database and provides several commercial products for electronic information retrieval. Initial exploratory contacts were made with the ICLR as a potential provider of legal information for this research but no reply was received.

5.5.2 The web scraper

It was necessary to design a system to traverse the BAILII website structure and to scrape data from the live systems. As has been mentioned previously, a design decision was taken to implement all non-text processing elements of the LARC application and preparation scripts in Ruby. Research highlighted that the *Anemone* gem would provide a solid basis for the work here.

Case reports on the BAILII site are organised under a hierarchy of index pages. The first level is ordered alphabetically in slices between different case names - e.g. *Barber R v...Barrier Ltd v* is one index page which includes all cases in that slice of the alphabet. The first step of LARC development was therefore to create a spider which would collect all the sliced index pages in the complete alphabetic list. See the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/> for the full code for the high-level index spider. The high-level indexes are listed on a complete A-Z directory web page and so this is the top level for the spider to consider. It then traverses links one level down from this page which have URLs that match a particular regular expression. The slice pages are indexed on URLs which can be identified by the following regular

expression: `/indices/ew-cases-[0-9]+/`. This makes the spider crawl all numerically-terminated URLs that start with `ew-cases-` one level below the main index page.

The URLs for case index pages that are discovered under the procedure above are stored in the `case-indices` table in the main database. There are columns in this table to indicate the date and time when the index page was added to the table and a boolean flag which indicates whether a particular page has already been crawled. The script takes the HTML source of each alphabetical page at one level down and creates an MD5 hash of the content. This hash is also written to the database and, each time the index spider is run, the MD5 hash for every crawled index is compared to the stored hash. This allows the system to know whether it needs to process case links on a pre-existing index entry or not.

Once the `case_indices` database table has been populated or updated with new links to crawl, a separate program is run to collect the content of individual case reports. See the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/> for the full case report scraper script. This code traverses all links one level down from the alphabetical slice index pages which match the regular expression `\ /cases \ /`. This is because each case report HTML page is located in a `cases` directory on the BAILII web server. Thus there is a simple mechanism to ensure that the spider collects only content from case reports and not from pages reached by the many other informational links on a typical BAILII page. The function of the case report spider is simply to collect the complete HTML source code for each case report. This is encoded in UTF-8 and stored in the `case_pages` database table. Each captured HTML source document is stored along with the URL of the page and the crawl date. The final stage of the initial data preparation process is to clean the HTML code for each report and to extract identifying and diagnostic information from it. This process is described in more detail in Section 5.6.1.

The heart of the LARC database schema that is described in Figure 5.2 is the `case_pages` table. This is populated from case report pages collected in the `case_indices` table, as described previously. `case_pages` contains the raw HTML source for each ingested case report, as discussed in Section 5.5.2. This table has a unique integer ID column which is set to auto-increment. The ID value for each report from this table is used as a canonical identifier for case reports throughout the data model for the system as a whole. The ID column acts in a foreign key relationship as a consistency constraint between this table and many other tables in the data model. The next level of the system centres on the `case_data` table, which contains case report information after it has been cleaned and faceted. The refinement sequence which results in an entry being placed in this table is described in more detail in Section 5.6.1.

Another central store in the model is the `users` table. This forms the basis of the authentication system which manages access to LARC and the roles and responsibilities which are associated with different accounts. The requirement for users to sign up for an account on LARC is a design decision which enables many features on the system, including saved sessions, saved contexts, activity timelines and so on. Authentication and identity provision also enables the collaborative document editor which is at the heart of the system. New users can have either standard privileges or they may themselves be declared as administrators. Administration rights enable a user to manage the system at a low level through user interface elements that are hidden for other users.

The `citations_for_cases` table holds the various different legal citations for individual case reports that can be discovered from the BAILII records. There is a many-to-one relationship between this table and the `case_data` table because an individual case report can have multiple valid citations. This arrangement is replicated for `case_judges` as there is often more than one judge who presides over an individual legal case. A particularly important data store in the schema is `citations_tables`. This stores data about all the citations of other cases that can be found in a particular report. It is implemented to facilitate the citation visualisation which is described in more detail in Section 5.8. This visualisation is augmented with information abstracted from the `sentiment_cases` table, which in turn contains processed information derived from the `sentiment_sentences` table. Briefly, the sentiment tables enable a treatment index for each case which is cited in an individual case report. This is visualised in the citation layout as

coloured edges between a case of interest and the cases that are cited in that report. There is a separate training interface and toolchain for the sentiment classifier.

The `documents_table` contains details of collaborative documents which are created and stored on the LARC system. Each document belongs to a particular system user. Groups of users can work on any document on the platform at any time. The `views_tables` store contains coordinate positions for different panes in the research view interface. The main screen of LARC is implemented as a grid of panels which can be resized and moved around the screen, as can be seen in Figure 5.13. Different visual layouts can be saved as views which belong to a particular system user.

The `screenshots_tables` store contains information about captures of the citation diagram which are taken automatically on the server. This ability for the user to view historical citation diagrams that they have seen in their research forms a part of the saved contexts auditing tool which is described in more detail in Section 5.11. Finally, it is worth noting that the `legislation` tables in the entity-relationship diagram follow the same pattern of interdependency and constraint as the `case` tables. The only exception here is `legislation_subjects`, which contains processed information about the subjects of different legislative instruments from the TNA XML data. There is a many-to-one relationship between this table and `legislation_data` because one act can concern many different topics.

5.6.1 How to identify relevant content

For case reports, once the raw source of all documents has been loaded into the `case_pages` database table, it is then necessary to clean and systematise the data. This is important both to remove styling and formatting code that is specific to the BAILII site and to provide the underlying systematic data required for search facet searches in LARC. A preparation script called the `case_cleaner` integrates most of the functionality required to take source from the `case_pages` table and to transform it into one entry per case in the `case_data` database table - see the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>. The cleaner script loads the raw source code for each page which was stored in previous steps and uses an XML parser to isolate case title and primary citation data.

A waterfall method is used because there is some variability of formatting between different pages. Most reports, for example, use square brackets to enclose the date of a case in the page title. This is the standard legal notation. In some outlying instances, however, the date is enclosed in standard elliptical brackets.

Once the relevant identifying information for each case report has been processed and stored, the next step is to take the HTML source code for each full report and clean it. The intention here is to extract plain text for each report from the complicated HTML markup which is employed on the BAILII site. The Python script for cleaning this material can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

A significant amount of time was spent evaluating different plain text extraction algorithms that could reliably preserve the information content whilst discarding boilerplate code. Solutions such as `html2text` and `BeautifulSoup` were found to have low reliability, to leave too much irrelevant markup in the processed results and often to discard valuable textual information that was rightly part of the case report data. The performance of different libraries was compared by manually establishing the relative proportions of a random sample of cleaned documents which comprised report text and that which was boilerplate that should ideally have been discarded. The evaluation finally resulted in the application of the `BoilerPipe` library, which is a boilerplate removal system implemented in Java that has a Python interface. `BoilerPipe` is described in [98] and has been designed precisely because of the low levels of performance evidenced with other possible solutions. Various extraction algorithms can be applied through `BoilerPipe` and the `Canola` extractor was the most reliable and efficient of the available alternatives, as described in [149].

The preparation process for extracting plain text from legislation was different and more complicated than that for case reports. The same systematisation of the legislation data was undertaken initially and the XML source was written to the `legislation_pages` database table so that a pristine copy of the data was available at all stages of the preparation process. The XML sources from The National Archives repository had then to be built back into HTML documents. This is because LARC itself uses simple formatting to present the text of legislation (and case reports) to the system user. An XML parser was used to build suitable source documents from the data file. Once a simple HTML document had been built, this was written to the `legislation_data` database table in a

`cleaned_html` LONGTEXT column. See the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

The process for systematising and cleaning source material for inclusion in the LARC database fulfils three important design decisions. First, the data should be made reliable and predictable enough to enable faceted search on different parameters across both case law and legislation. Secondly, legislation and case report data should be kept separate both in the database and the user interface. This is done to reduce result set size and contamination in the platform as results are presented to users. Thirdly, the underlying markup used to present sources should be simple, easy to read and as clean as possible. This ensures that LARC is easy to use for a long period of time in reading and knowledge synthesis activities.

5.7 Enabling search facets

Once the underlying information for LARC had been systematised and identifying material was codified in the database (as described in Section 5.6.1), it was necessary to expose these facets to the user so that searches across different properties could be undertaken. A design decision was taken early in the prototyping process that the user should be guided by the system as much as possible throughout the search process. Another decision was that searching through different facets should be separated in order to keep the user interface as simple and clean as possible. Too many existing solutions either use unguided search, which means that input may or may not produce any search results when it is submitted, or they allow for faceted search across different metadata all on one screen, which leads to a crowded and complicated user interface. The initial search interface for LARC can be seen in Figure 5.3 with the different types of search spread across a top menu and each page dedicated to interrogating the database on a single property.

The decision to guide the user in their search activities resulted in a focus on implementing auto-suggestion facilities in the user interface. This means that, as the user starts typing a case name to search for, for example, the system highlights possible full case names from the input as it builds in the text entry box. The case name autocomplete can be seen in Figure 5.4. Searching by case dates works differently. Here, the user selects a range of years that they are interested in. The system populates a dynamic table on screen with cases that fall in the correct

range. Individual cases from the table can be isolated through a textual search facility which covers all fields in the table and is available in the table header. See Figure 5.5 for a date view populated through an initial date range.

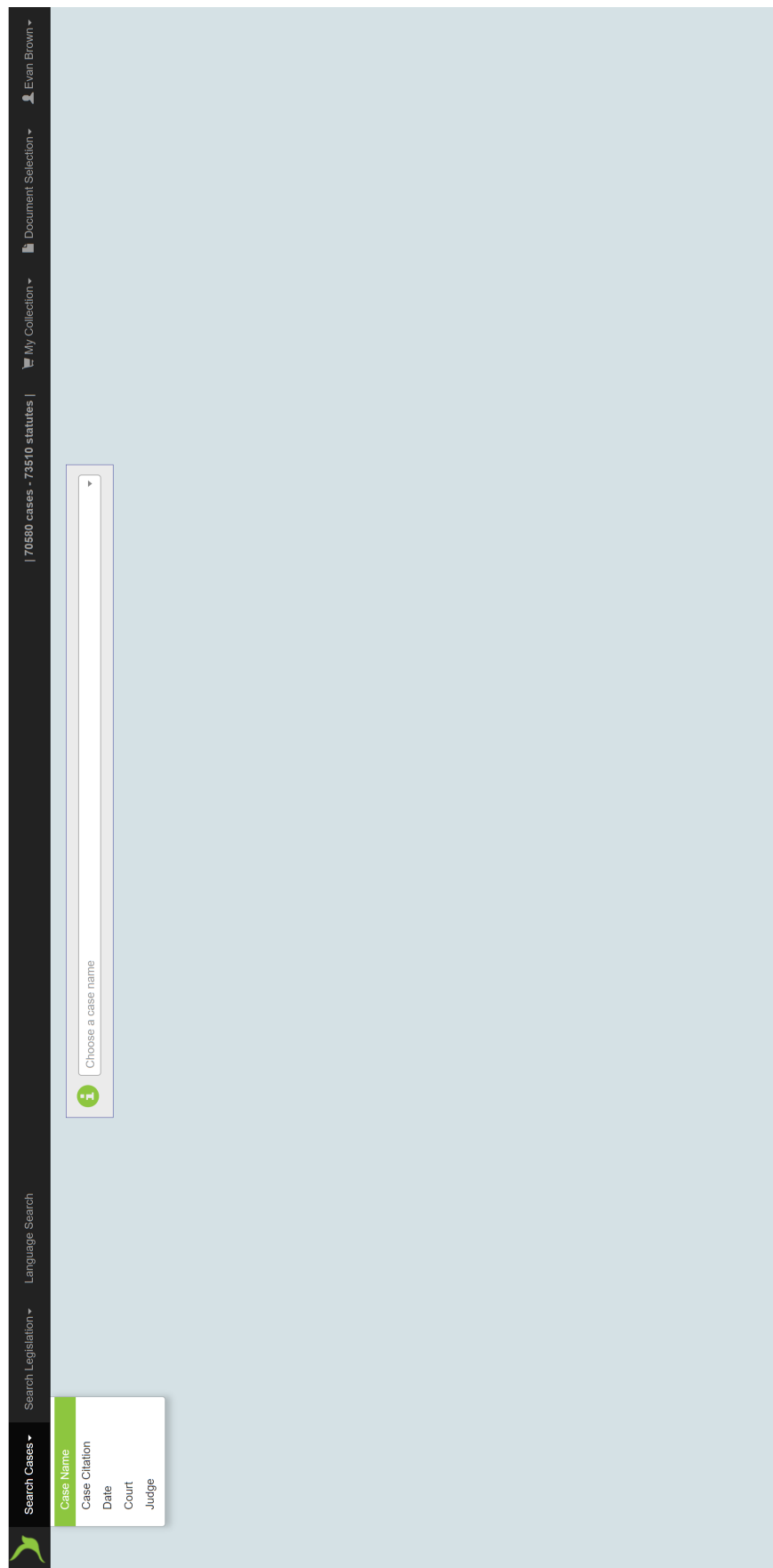
The facility to search by case judge employs a waterfall search methodology which allows the user to select an initial judge of interest through the standard auto-suggestion facility. The system then identifies all cases that the selected judge presided over. It allows the user to progressively focus their search by choosing additional judges that have sat with the initially-selected judge on different cases. The user can come out of the waterfall at any time by selecting the *Alone - with any other judge* option, and the appropriate cases are then used to populate a table on screen. This search facet is shown in Figure 5.6.

The auto-suggestion interface for searches implements a citation index which is displayed as a series of yellow stars in the text input box for each suggested case. The citation index displays a different number of stars, up to a total of five, depending on how often a case has been cited in other cases in the database. This facility has been implemented in a data preparation step which counts the number of times a case is cited in all the case reports in the database. These numeric measures are then quantised into buckets between zero and five in order to enable the star rating system. See the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/> for the details of the implementation here. The use of a citation index represents a design decision that the relative importance and weight of different cases should be communicated intuitively to the system user throughout their search activities.

The backend functionality of the auto-suggestion facilities was initially implemented as a web service which used SQL `LIKE` queries to progressively find case names and other data that matched user input. This proved unsatisfactory for two reasons. Firstly, repeated queries against the full database were relatively slow, with an average response time of 30ms. Secondly, the first implementation only suggested results in a forwards direction from user input. This meant that a famous case like *Bolton v Stone* was not found because it was indexed in LARC as *Stone v Bolton*. The auto suggest took the letter *B* at the beginning of the input string and only suggested completions that started with *B*. To remedy both of these issues, the auto-suggestion backend was moved to a Redis data store in memory on the LARC server which can implement search querying anywhere in a key string. The average response time of the auto suggestion mechanism was

cut to 3ms per input character as a result.

5. LARC - THE LEGAL RESEARCH AND COLLABORATION PLATFORM



150 **Figure 5.3:** The initial search interface in LARC with case facet options displayed.

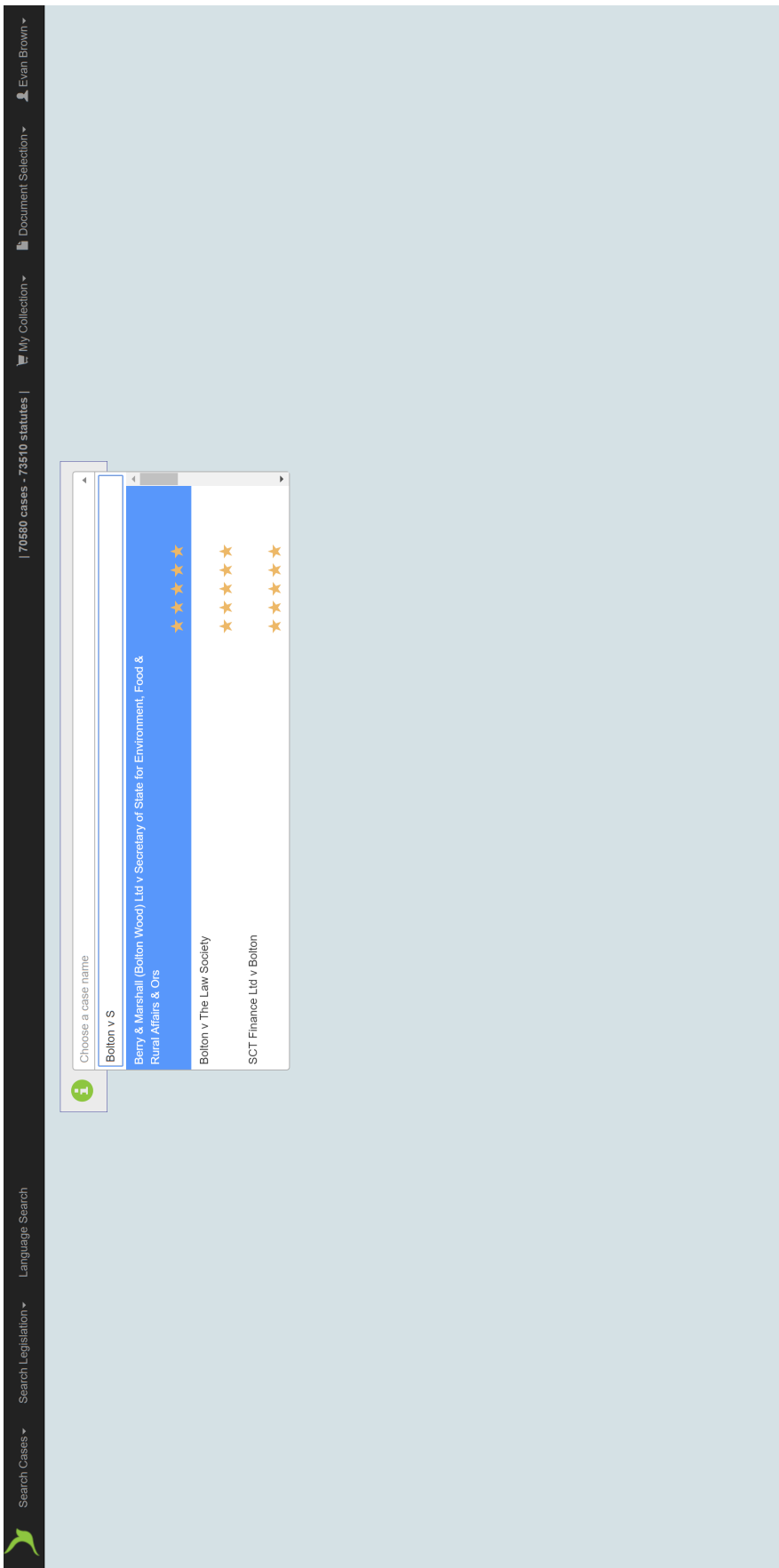


Figure 5.4: The case name auto-suggestion system in the LARC search interface 151

Figure 5.5: The date search facet in the LARC interface

152

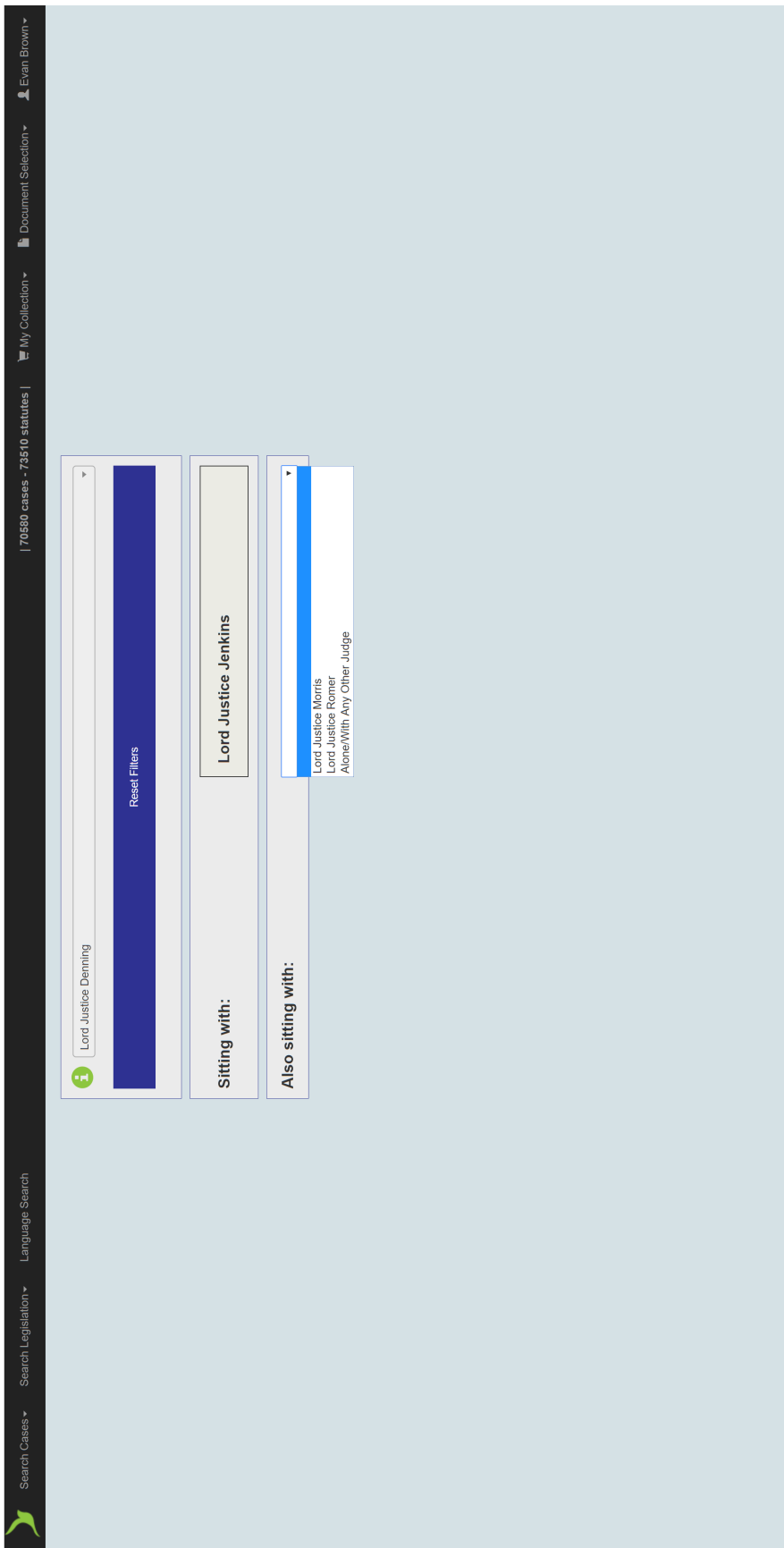


Figure 5.6: The judge search facet - demonstrating the waterfall - in the LARC interface.

5. LARC - THE LEGAL RESEARCH AND COLLABORATION PLATFORM

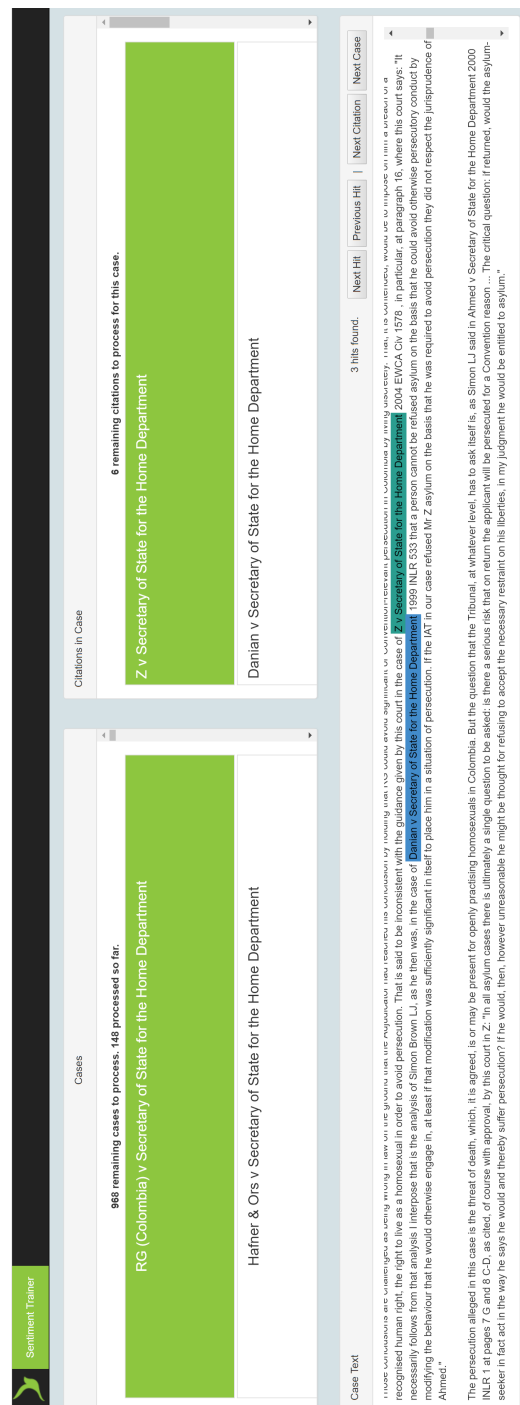


Figure 5.7: The sentiment trainer user interface

5.8 Identifying case connections and visualising them

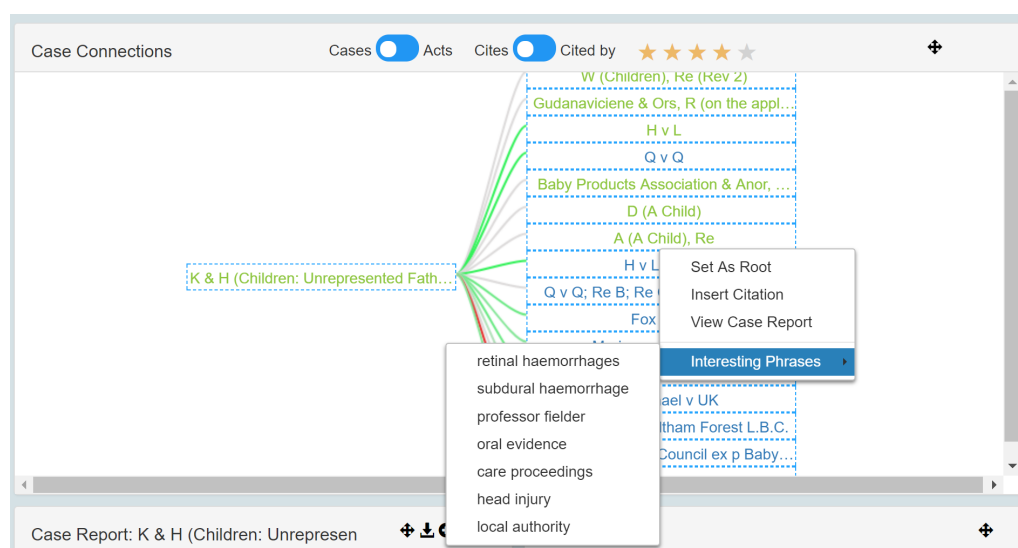


Figure 5.8: Citation visualisation.

Figure 5.8 shows the citation visualisation in Research View. This is a tree diagram which displays those cases that are cited in the currently-selected case. The colour of the edges in the diagram show how each cited case has been treated by the judge in the present case. There are toggles to switch between cited cases and cited acts, and also cases or acts cited in the case or cases that cite the current case. The tree implements on-demand node loading. This means that the user can click on any case in the tree to expand the diagram to show the cases cited in that case. Case names rendered in green are available as full entries with report text in the LARC database. Blue entries are citations for which no data is available.

Edge colours are based on automatic sentiment analysis of the text around citations in the case reports. Sentiment scores are attached to each citation in the database in advance through a pre-processing script. The scores are generated through a classifier which has been trained manually on language around a collection of 10,000 citations from the database. Green edges represent positive treatment, a neutral treatment is grey and negative sentiment is shown in red. Positive treatment broadly equates to a linguistic context where a principle or line of reasoning from a cited case has been applied in the present circumstances, or at least referred to favourably by the judge. Negative treatment equates broadly to a

context where a finding or line of reasoning in a cited case has been differentiated from or over-ridden in the current case. Scores are based on the number of positive n-grams around a citation minus the number of negative n-grams. The sentiment scores are processed by a quantile cutting algorithm to bin them into twelve steps on a gradient between green and red, through grey. Thus the strength of colour in an edge indicates how strong the sentiment associated with a citation is.

The sentiment visualisation is trained through a sub-application in the LARC platform code. The user interface for this program can be seen in Figure 5.7. The system selects 10,000 citations from different case reports at random when the classifier is initially in an untrained state. In the top left of the sentiment interface, those cases which have been mined for citations are displayed in order. In the top right, each citation for the current case is shown. The system automatically moves through every citation in a case, and then through the cases as the available citations are exhausted. The case report text is shown at the bottom of the screen and each citation is highlighted within the report. The training user reads the text which occurs around a citation, right-clicks to highlight it and then identifies whether the selected language is positive or negative from the resultant menu. There are controls which allow the user to jump forwards and backwards through the case report as they conduct their work. The system automatically scrolls to each following citation in a report once a treatment has been selected.

The sentences that are highlighted and classified here are written to the `sentiment_sentences` database table, together with a boolean flag which identifies whether the language is positive or negative in nature. Once the ten thousand citations have been dealt with manually in this manner, it is then possible to train a sentiment classifier using the positive and negative sentences as input. The sentences are loaded from the database into two lists, one which contains positive language and the other negative. A `NaiveBayesClassifier` from the Natural Language ToolKit for Python is used here. It is fed a tuple for each sentence which contains the text and a designation of either `pos` or `neg`. An associated position element is also written. This codifies where the positive or negative language was found in relation to the citation as a character offset. The code for this element of the process can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

The classifier is saved to disk as a pickled binary object so that it can be used in further stages of the classification process. Next, the classifier which has been

previously trained is used to categorise all the citations in the complete collection of case reports. The position attributes which were saved in the previous step are averaged for both positive and negative. This helps to ascertain how much language around each citation must be fed into the classifier in order to be likely to contain important language. This element of program code can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

The classifier breaks the positive and negative sentences from the training set down into n-grams and identifies occurrences of these collocations around the novel citations. Ultimately, an aggregate score for each citation is computed which is derived from the number of positive n-grams found minus the number of negative n-grams found. An overall score of more than zero is treated as positive, less than zero as negative and zero as neutral. In this scheme, citations which cannot be classified because the surrounding language is neither categorised as positive or negative default to a zero score and a neutral classification.

In the final step, a separate preparation script iterates through the database of case reports and processes each sentiment score that was derived previously. These are quantised into twelve bins which correspond to steps on a gradient from red, through grey to green. Each bin in the sequence is associated with a hex colour which represents its position in the gradient. Every sentiment score in the database is assigned a colour which is then written to the database. In this manner, the treatment of a given citation which is expressed in the overall sentiment score is converted to an edge colour for display in the citation visualisation.

5.9 Supporting note taking

The heart of the LARC user interface is a collaborative document editor which supports synchronous collaboration. Once a user has searched for an initial case or item of legislation, the system asks them to select an existing document to work on or to create a new one. Once a document has been chosen, they are switched from the search interface to *Research View*, which can be seen in Figure 5.13. By default, the document editor loads the selected document into a pane on the left hand side of the interface. The initial idea is to keep the document in a long vertical pane for ease of use and layout. When a new user is registered on the LARC system, they are assigned a unique authorship colour. All text entered by an individual user is

highlighted with their authorship colour in the background.

The document editor is based on a highly-customised version of the open source *EtherPad* project. The selected document updates for all users who are viewing and working on it in real time as other people make changes. The system supports up to 16 people all working on a single document at the same time. LARC documents are implemented as group pads in the Etherpad system. A session variable is set and read by the software on invocation which authorises a particular user to work on a given pad. This authentication and permissions system also means that authorship colours persist throughout the life of a pad, so text which has previously been entered by a given user in a prior session retains their correct authorship colour. The mechanics of the authorship system and the authentication procedure can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

LARC allows the user to switch document or to start work on a new document without changing the existing state of the research view. Indeed, new cases and items of legislation can also be searched for without changing the document. Switching document changes the document panel and the chatroom (because chatrooms are tied to individual documents) but leaves the citation visualisation and the case report undisturbed. Searching for a new case or an item of legislation changes the citation visualisation and the case report text panel without interfering with the document or the current chatroom. This reflects a design decision to minimise context switching as the process of legal research proceeds and develops.

5.10 Instant messaging and chat facilities

The LARC platform includes an instant messaging facility which allows users to talk to each other whilst they work on legal research tasks. A chatroom is created for each document on the system whenever a new pad is invoked. This means that everybody who is working on a particular document at one time can communicate with each other directly from research view. The chatroom for every document and all historical message content persists between sessions until a document is explicitly deleted by an administrator. The chat facility is based on a heavily customised version of the open source *converse.js* project.

Research view is an integrated environment and, as such, case citations and the text from case reports can be automatically inserted into a chat message by right-

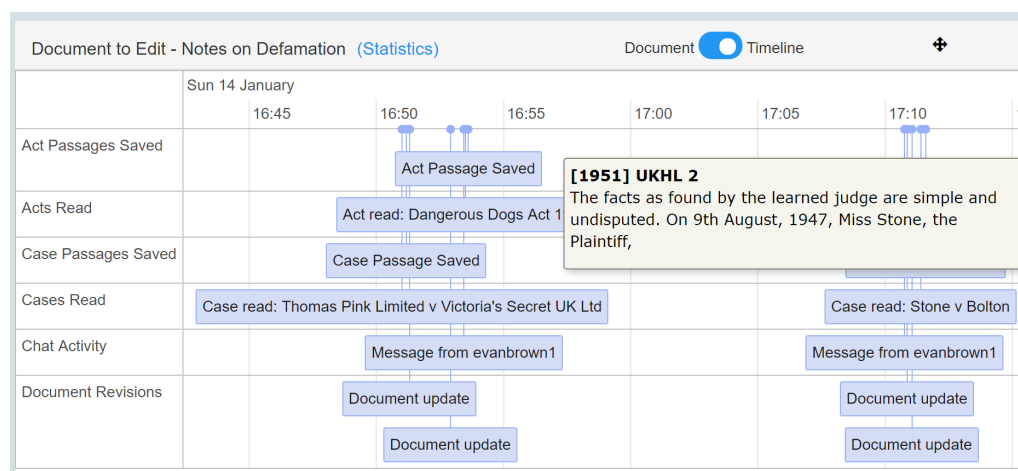


Figure 5.9: Integrated Timeline Visualisation

clicking the content and selecting *Insert into chat* from the resultant contextual menu. This is a design decision which seeks to simplify the process of sharing cases and legislation that has been discovered by one user amongst the entire group which is working on a particular document.

5.11 Auditing tools

LARC provides a timeline visualisation which draws data from the different components of the system into a single account of activity related to a particular document (Figure 5.9). This is designed to facilitate the auditing of progress and the direction of research in a team by tutors or senior lawyers. The timeline is delivered by *vis.js*. This has been extended so that information about the content of different data points is displayed on hovering the mouse over an entry in the visualisation. The timeline is navigable by scrolling the mouse both in the horizontal and the vertical. The collaborative document editor saves document revision information in deltas to the MySQL database. These revisions are parsed for display in the timeline so that a new revision entry is shown when a paragraph of text has been added to the document or removed from it. Session data for saved cases, acts and textual passages form parts of the visualisation, together with chat message content from all the users who have worked on the current document. The timeline visualisation has another view which is accessible from within the collaborative document itself. A slider allows users to move between document revisions. Content is added and removed from the screen as appropriate using an

animated transition. This offers a granular assessment of progress.

The timeline utilises a customised XML export functionality in both the *Etherpad* document editor and the *converse.js* chat client to pull histories of document revisions and chat messages into the timeline. These data include timestamps and chat or document revision content. *Etherpad* implements a web service for XML request and response whereas *converse.js* writes the developing XML data of chat content to the main LARC database in a dedicated table. The timeline creation program considers all waypoints in the document but only displays a revision with contextual text display when a significant number of sentences has been manipulated in the editor. The controller action which controls this behaviour can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

Saved Research Context

Document content

In contracts for the **sale of goods** and **supply of services** certain basic provisions are implied by statute in order to provide protection to purchasers. In consumer contracts, the provisions derive from the **Consumer Rights Act 2015**. The **Consumer Rights Act 2015** came into force on 1st Oct 2015 and replaced many of the provisions contained in the **Sale of Goods Act 1979** and the **Supply of Goods and Services Act 1982** where there is a consumer sale. The **Sale of Goods Act 1979** and the **Supply of Goods and Services Act 1982** have not been repealed and still apply to contracts for the sale of goods and the supply of services outside a consumer context (eg private sales and business to business transactions).

In addition there are implied terms that the service must be carried out with reasonable care and skill, that the service will be carried out within a reasonable time and where no price is agreed a reasonable price will be paid. These protections are in the form of statutory implied terms. This means that the **Consumer Rights Act** or the **Sale of Goods Act** will put these terms into all contracts for the sale of goods no matter what the parties themselves have agreed in the terms and conditions of sale.

Harrington and Leinster Enterprises Ltd v Christopher Hull Fine Art Ltd [1989] EWCA Civ 4

The main protection offered covers where the seller does not have the right to sell the goods, where the goods are sold by description there is an implied term that the goods will correspond to that description, businesses must ensure that the goods they sell are of satisfactory quality and fit for their purpose, where the goods are sold by sample there is an implied term that the goods will correspond to the sample in quality.

Here is some new text: Page UK Ltd & Anor v Chobani UK Ltd & Anor [2014] EWCA Civ 5

Precedent diagram - Click image to launch viewer

East v Maurer

Allied Maples Group Ltd v Simmonds &

Stone Heritage Developments Ltd & Or

British Transport Commission v Gourdes

Aerospace Publishing Ltd v Thames W

Livingstone v Rawyards Coal Co

Smith New Court Securities Ltd v Citib

in East v Maurer

Cullinane v British "Hera" Manufactur

Jenmain Builders Ltd v Steed & Steed

Allied Maples Ltd v Simmonds & Simm

Stone Heritage Developments Ltd v Da

4 Eng Ltd v Harper & Anor

Case report

James v Thomas - [2007] EWCA Civ 1212

Miss Joanne Wicks (instructed by Turnbull Gairard) for the Appellant Sir John Chadwick :

This is an appeal from an order made on 6 July 2006 by His Honour Judge McKenna sitting in the Birmingham County Court in

Chat history

morrisandnews36

2018-06-28 15:12:24 +0100

Case citation: Harrington and Leinster Enterprises Ltd v Christopher Hull Fine Art Ltd [1989] EWCA Civ 4

evanbrown26

Occupants

● evanbrown26

Context Saved

Context Saved

Context Saved

Clive Freedman QC and Ian Smith (instructed by Reid Minty LLP)

for the Claimant Tom Leech (instructed by Daniel Berman & Co)

for the First Defendant Nigel Hood (instructed by Byrne & Partners) for the Second Defendant The Hon. Mr Justice David Richards:

Summary judgment was entered in this action on 3 May 2007.

Figure 5.10: The saved context view in the LARC interface.

5.11.1 Saved contexts

The timeline functionality is augmented with a facility for viewing saved contexts, as shown in Figure 5.10. This means that the software saves a waypoint for the entire research view interface when a case or statute is selected or changed or the document being edited is substantially altered. There is a group of entries within the timeline visualisation for saved contexts. When any of these entries are clicked on, a modal dialogue opens with a facsimile of the research view interface in it. The content of the interface copy represents the case report or statute, the document content, the chat content and the citation diagram as it was when the waypoint was saved. Thus it is possible for users and for senior lawyers or teachers to step back in time in a research activity and to understand how progress occurred and how research directions developed over time.

The citation diagram is technically composed of an HTML5 canvas object with vector paths for the edges. The case names and act names are enclosed in DIVs which are located precisely over the canvas to produce an integrated visualisation. A screenshot entry with the file path and a timestamp is written to the database in the `screenshots_tables` store. This means that captures of the citation diagram are taken entirely and transparently in the backend infrastructure but they mirror exactly what an individual user has been seeing at a particular point in time. The controller action which enables saved contexts can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

5.11.2 Statistics

The LARC interface provides a statistics feature which collates and presents information about the document that is currently being worked on. Information in this view includes the users who have contributed to the document; their authorship colours; the number of words and characters that each person has written; and the number of whole sentences and paragraphs that each user is responsible for writing. This feature reflects a design decision to provide simple tools which allow senior lawyers or teachers to ascertain who in a group is responsible for dominant research directions, who the principal contributors to finished work products are and which of the contributors may have struggled with a particular legal research activity. The statistics dialogue can be seen in Figure 5.11 and the underlying program code can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

5. LARC - THE LEGAL RESEARCH AND COLLABORATION PLATFORM

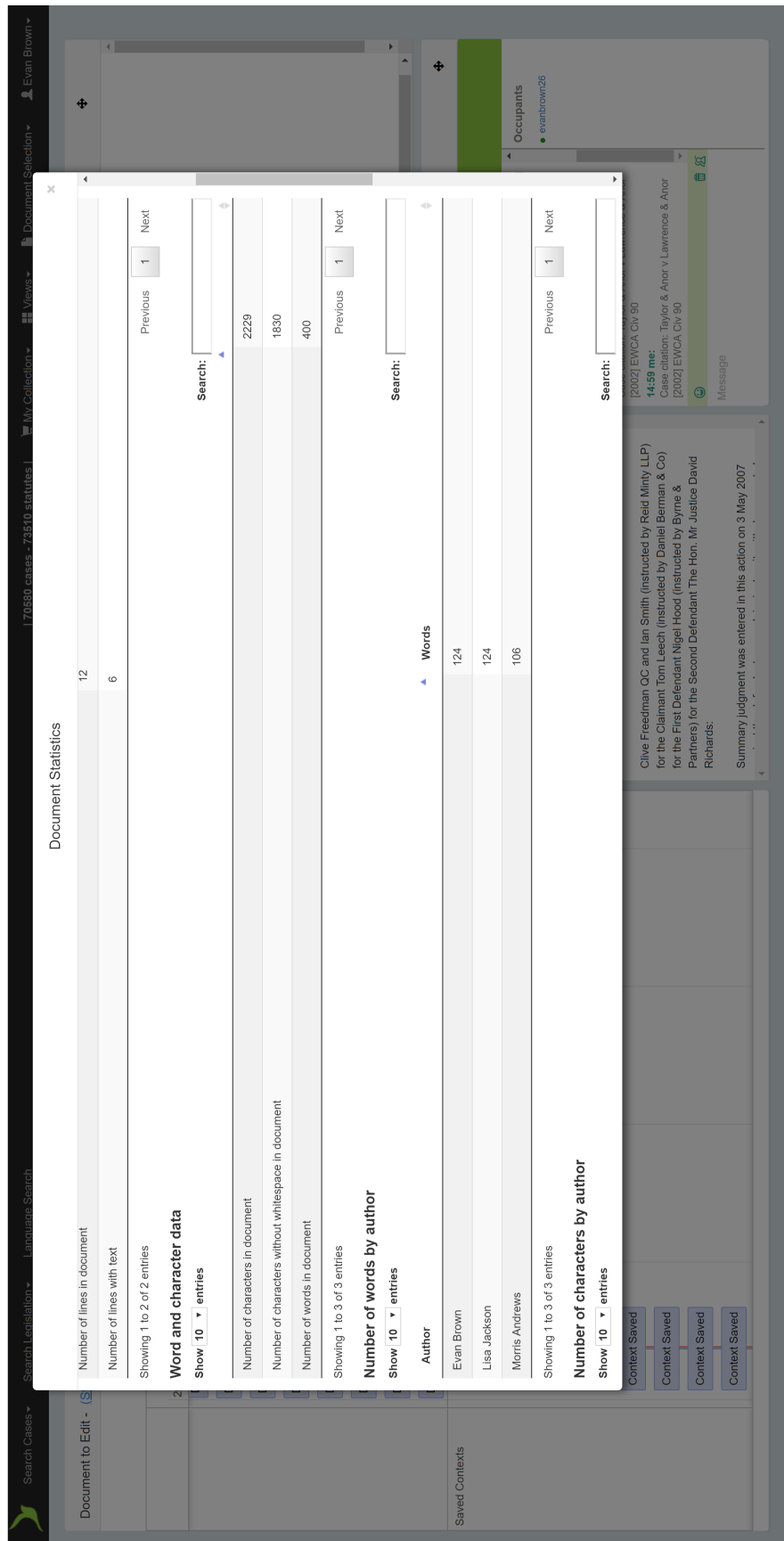


Figure 5.11: Document statistics for the current pad in LARC.

The screenshot displays the LARC sessions manager interface. The top navigation bar includes links for 'Search Cases', 'Search Legislation', 'Language Search', '70580 cases - 73510 statutes', 'My Collection', 'Document Selection', and a user profile for 'Evan Brown'. The main content area is titled 'Session History' and is divided into two panels.

The left panel, titled 'Sessions', contains a table of session records:

Sessions
2018-03-27 14:34:36 UTC
2018-04-24 13:48:19 UTC
2018-04-30 09:06:56 UTC
2018-04-30 11:34:41 UTC
2018-06-14 13:31:16 UTC
2018-06-29 14:02:46 UTC
2018-06-30 11:35:09 UTC

The right panel, titled 'Saved Cases', shows a detailed view of a case. It includes a search bar, a 'Show 10 entries' dropdown, and a table with the following data:

Case ID	Case Title	Case Citation	Link	Delete
59881	Thomas Pink Limited v Victoria's Secret UK Ltd	[2014] EWHC 2631	View	Remove

Below the table, it indicates 'Showing 1 to 1 of 1 entries'. At the bottom of the panel, there are 'Previous' and 'Next' navigation buttons.

At the bottom of the interface, there are links for 'Delete Session' and 'Export Session Data'.

Figure 5.12: The sessions manager for the shopping cart in LARC.

5.11.3 Sessions and history

Figure 5.12 shows the management interface in LARC for sessions and saved data. The system implements a shopping basket which belongs to each individual user of the platform. When they authenticate and log in, a new session record is created in the `sessions_tables` database store. These sessions are recorded and saved until the user or a system administrator explicitly deletes them. Each shopping basket persists for one login session before being archived. A user can navigate backwards through their saved sessions in the management screen which is shown in the figure above.

The user can save a broad range of data to the shopping basket, This includes cases and statutes with their appropriate citations, text passages of interest from relevant cases and acts, and elements of the linguistic search process which will be described in detail in Chapter 6. Adding information to the session is accomplished easily through the research view interface in Figure 5.13. There is a folder icon in the top of the case report and act content pane which saves the current case or statute to the session when it is clicked. Textual content from the current report can be saved by highlighting it in the same pane and choosing *Add text to collection* from the contextual menu. The implementation of sessions in LARC reflects a design decision that users should be able to “bookmark” important information that they discover in a persistent and easily navigable manner.

5.11.4 Interesting phrases

All the case reports in the database are processed in advance for frequent noun phrases. This allows for the interesting phrases menu which is shown in Figure 5.8. Phrases are intended to be diagnostic of the dominant themes in a case and their identification is based upon work by Pala et al in [144]. The phrase display is used to augment the citation visualisation so that a user can click on a case of interest in the diagram and call up a list of the most commonly-occurring n-grams which are important in that case report. The on-demand node loading in this visualisation means that, as well as phrases in the root case, any linked case is automatically parsed for important phrases as the diagram is extended. Clicking on an interesting phrase in the contextual menu launches a linguistic search across the entire corpus of legal text, or a more limited evaluation for that language in cases and acts contained in the shopping basket at the time of the request. The details and implementation of language search in LARC will be explored in detail

in Chapter 6.

The phrase extractor is a preparation script which runs on case reports and statute text in advance. Data is stored in the `case_data` table as JSON which is parsed on demand by the front-end visualisation code. The script iterates over the plain text report content for each case or statute and initially uses a noun-phrase extractor to develop candidate terms for inclusion in the visualisation. This extraction procedure is enabled by the attribution of Part Of Speech (POS) tags to every word in the case report text before it is processed further. A number of important steps are then taken to ensure that only high quality information is saved as phrases to the database.

First, n-grams which feature proper nouns (names) are excluded. This is because cases frequently feature language around particular people in the action. These high frequency phrases are of limited use. Secondly, a phrase is only added to the collection if it has not been seen before in the case being processed. This ensures the elimination of duplication in the *Interesting phrases* menu. Thirdly, a Python interface to the *WordNet* database is used to check that each constituent word of a phrase appears in the dictionary. This is done to guard against dirty data in the case report with mis-spelled, ungrammatical or partial content being excluded. Finally, the noun-phrases for a case are organised in descending order of frequency and stored. This means that phrases which appear at the top of the contextual menu in the citation diagram occur more often than those which are placed lower down. The script for noun-phrase extraction can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

5. LARC - THE LEGAL RESEARCH AND COLLABORATION PLATFORM

The screenshot displays the LARC Research View Interface, a web-based platform for legal research and collaboration. The interface is divided into several main sections:

- Top Bar:** Includes navigation links for "Search Cases", "Search Legislation", and "Language Search". It also shows the user's profile "Evan Brown" and a "Document Selection" menu.
- Document Editor (Left Panel):** Features a "Document" tab with a timeline view (0 to 22). The main text area displays a document titled "Trademark Infringement" with a large letter 'A' in a circle. The text discusses a trademark infringement case involving "Thomas Pink Limited v Victoria's Secret UK Ltd [2014] EWHC 2631".
- Case Connections (Middle Panel):** A central area showing a network of cases. A large letter 'B' in a circle is prominent. Cases listed include "Total Ltd v YouView TV Ltd", "Stichting BDO & Ors v BDO Unibank", "Specavers International Healthcare Ltd v BDO Unibank", "YouView TV Ltd v Total Ltd", "Maler & Anor v Asos Plc & Anor", "West Country Renovations Ltd v Mc...", "Comic Enterprises Ltd v Twentieth C...", "Jack Wills Ltd v House of Fraser (St...", "YouView TV v Total", "Pan World Brands v Tripp", "Roger Maler v Asos Plc", "Starbucks v British Sky Broadcasting", "Coliseum Holding AG v Levi Straus", "Comic Enterprises v Twentieth Cent...", and "Interfora v Marks & Spencer".
- Case Report (Right Panel):** Displays a detailed report for "Thomas Pink Limited v Victoria's Secret UK Ltd [2014] EWHC 2631". The report includes a "Search" section with "0 hits found" and a "Primary citation" section. The "Before" section lists the parties: "Charlotte May QC and Asa J. Forster, instructed by Bristows LLP" for the Claimant and "Mr Justice Bliss" for the Defendant. The "After" section lists the parties: "The claimant, Thomas Pink Limited, a company trading in London in 1984. The core of its business is and has always been the sale of shirts which are worn by professional people, particularly men. Its flagship store is on Jermyn Street in London. Jermyn Street is famous for its shirt makers." and "The defendant, Mr Justice Bliss".
- Chat (Bottom Right Panel):** A chat window titled "Chat" with a "Message" input field. It shows a conversation between "dundee007" and "bernardjackson3". The chat history includes messages from "bernardjackson3" and "dundee007".

Figure 5.13: The LARC Research View Interface.

Home

New to JustCite? Register

Sign In

JustCite

The good law guide

donoghue

Advanced search

1 / 3

Print

Email

Download

Sort by: JustCite Ranking

Showing 1-25 of 70 results for donoghue

Everything (70)

Cases (63)

Journal Articles (3)

Barristers & Chambers (1)

Text filter

Filter by jurisdiction

M'Alister or Donoghue (Pauper) v. Stevenson. - 1932

116 positive

975 neutral

17 negative

Recent: Dicta of Lord Atkin applied by Attorney General of the British Virgin Islands v H (PC)

[1932] UKHL J0526-1

CASE

UK

State (Healy) v Donoghue - 1976

21 positive

194 neutral

4 negative

Recent: Referred to by (2011) 52 EHRR 20 (ECHR)

CONSTITUTION - Courts - Administration of Justice

[1976] IR 325

CASE

IE

Poplar Housing and Regeneration Community Association Ltd v Donoghue - 2001

Housing - Assured shorthold tenancy - Order for possession

1 positive

49 neutral

1 negative

Recent: Referred to by DPP v Lukaszewicz (CA (Ire))

[2002] QB 48

CASE

UK

Patrick Donoghue v DPP - 2014

CRIMINAL LAW - Judicial review - Prosecutorial delay

2 positive

10 neutral

Recent: Referred to by DPP v Lukaszewicz (CA (Ire))

[2014] 2 IR 762

CASE

IE

Donoghue v Burke and Another. - 1960

TORT - Joint tortfeasors - Apportionment of liability

2 positive

2 neutral

Recent: Applied by Snell v Haughton (Supreme Court (Ire))

[1960] IR 314

CASE

IE

Donoghue v Allied Newspapers Ltd - 1938

Intellectual property - Copyright - Author

5 neutral

Recent: Referred to by [1971] HCA 48 (High Court (Australia))

[1938] Ch 106

CASE

UK

THE KING (KATE DONOGHUE AND OTHERS) V THE JUSTICES OF COUNTY CORK - 1910

JUSTICES OF THE PEACE. - Bias - Personal ill-will towards defendants

3 neutral

1 negative

Recent: Considered by [1996] 1 LRC 342 (Supreme Court (India))

[1910] 2 IR 271

CASE

IE

R v Donoghue - 1987

(1987) 86 CrAppR 267

Figure 5.14: The case-centric search interface in JustCite.

5. LARC - THE LEGAL RESEARCH AND COLLABORATION PLATFORM

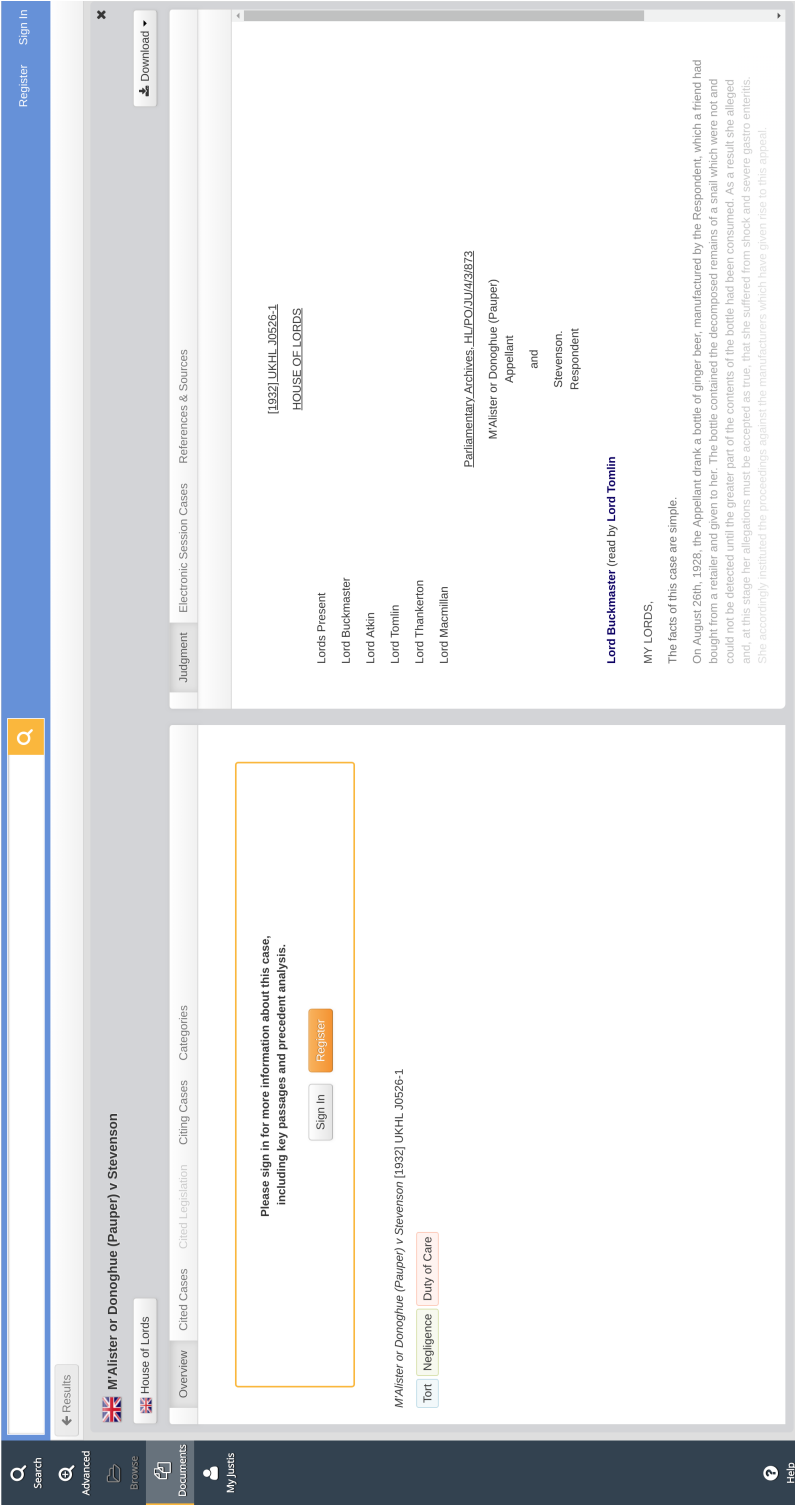


Figure 5.15: The case reading interface in JustCite.

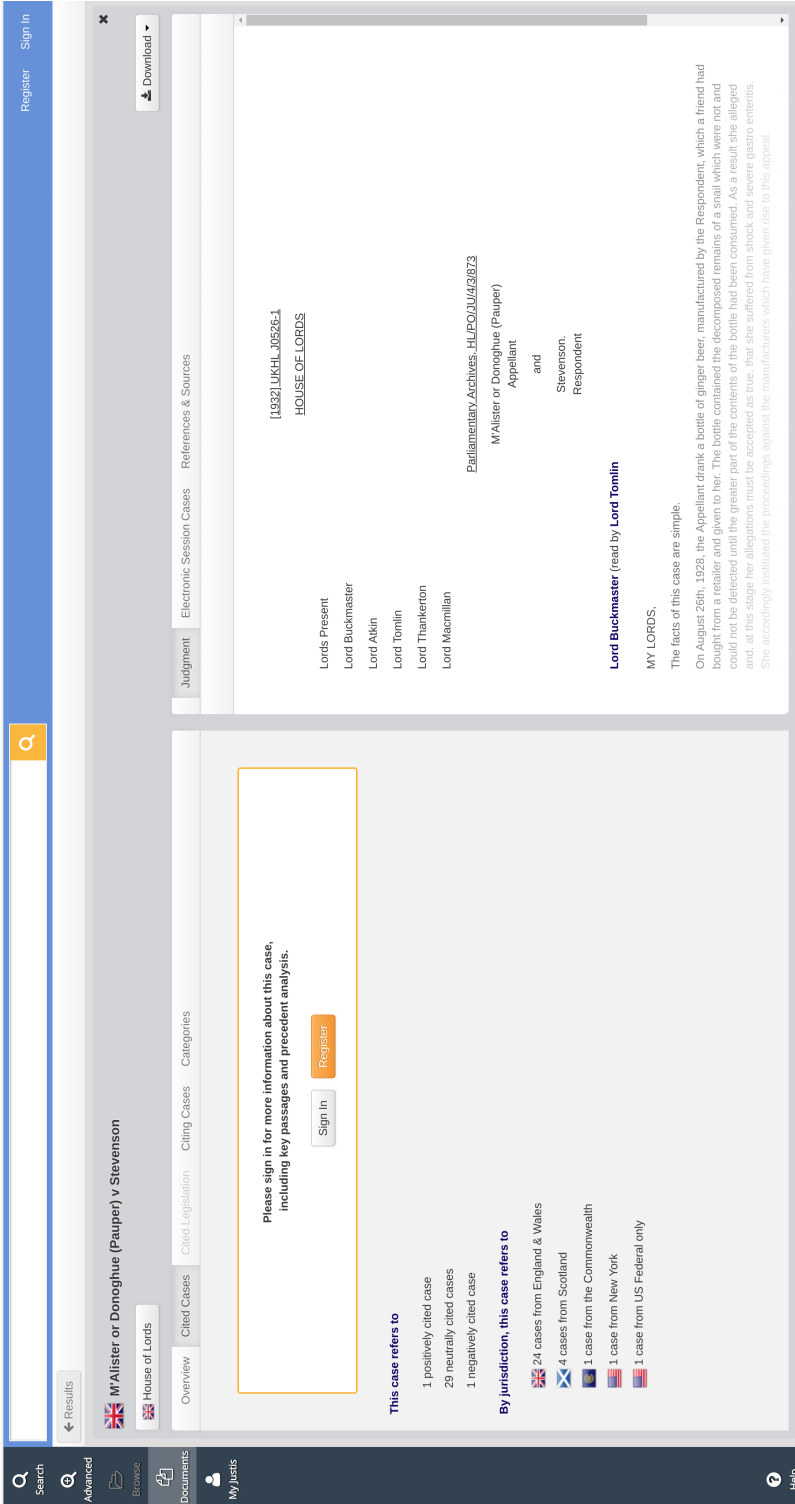
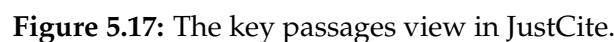


Figure 5.16: The cited cases view in JustCite

172



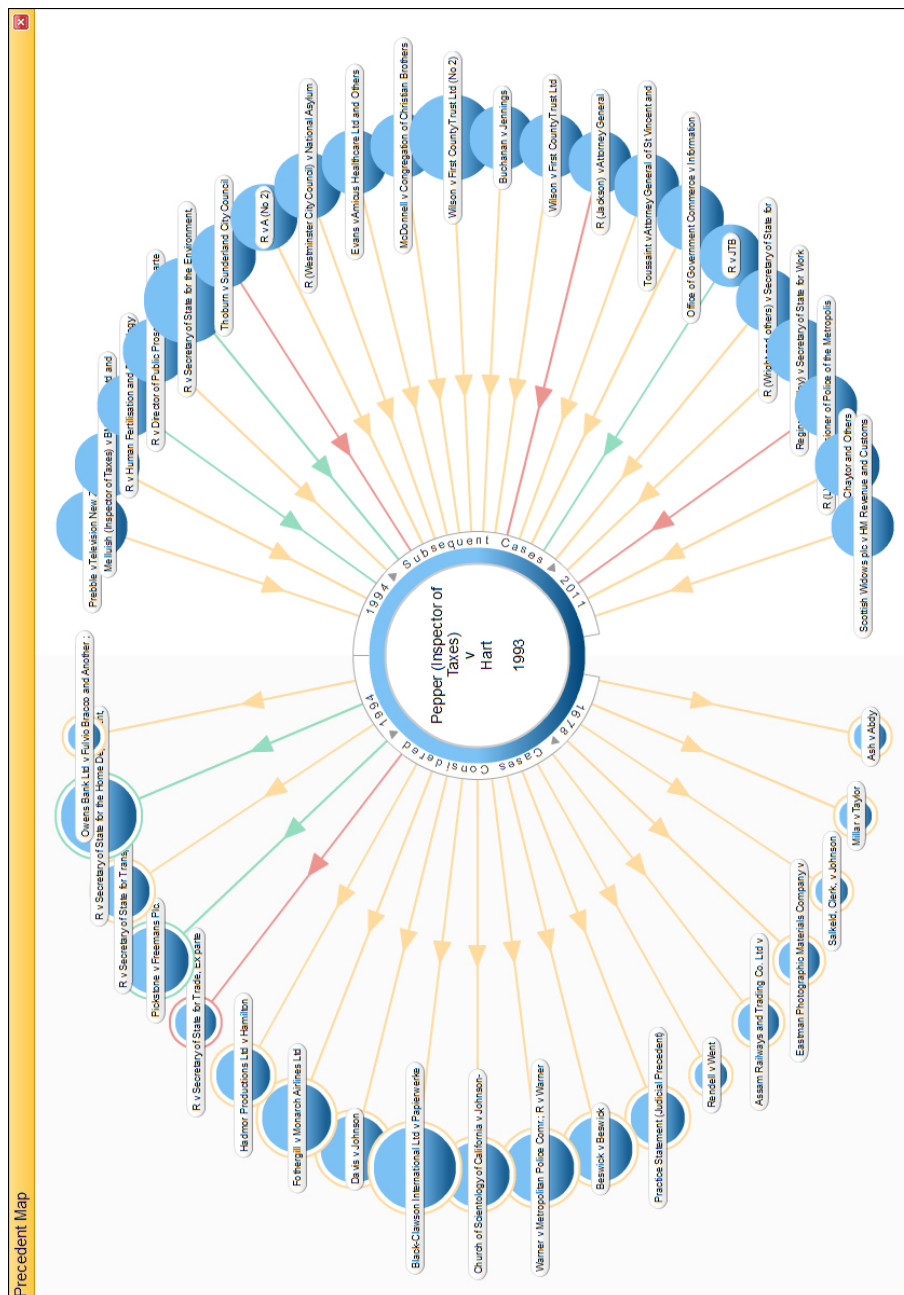


Figure 5.18: The precedent diagram in JustCite.

5.12 Summary: Bringing it all together - the complete LARC user interface

The default interface in LARC is called *Research View*, as can be seen in Figure 5.13. The user initially selects a case or item of legislation as a seed for their work. They can search the database across multiple facets to find this initial instrument. This flow was based upon our observation of moot activity. Students are given a number of cases and items of legislation initially to guide their preparations. Once an initial case or act has been selected, the system switches to *Research View*. The Research View screen is divided into four panels. The position of each panel is configurable by drag-and-drop. Panel configurations can be saved as views on the system and the user can set a default view or switch between different views at any time. A lack of configuration options for look and feel in current tools was the impetus for implementing this functionality in LARC. The four screen panels contain the collaborative document editor (marked **A** on Figure 5.13 - see Section 5.9), the citation map for cases and acts (marked **B** on Figure 5.13 - see Section 5.8), the text of the case report or act itself (marked **C** on Figure 5.13 - see Section 5.6.1) and the chat window (marked **D** on Figure 5.13 - see Section 5.10). The document to work on is created or selected by the user after they have chosen an initial case or statute. A chatroom is created or recalled for each document on the fly. Users can switch between documents and associated chatrooms using a drop-down menu in the page header without losing their place in the current case.

Every time a user authenticates with the LARC system, a session is created and stored to the MySQL database. The user can save cases and acts that they find of interest to a basket (called *My Collection*) for later recall. They can also store particular passages from instruments to their basket by highlighting text in the report pane and clicking the *Save To Collection* option in the contextual menu. There is a management interface for sessions which is accessible from the top menu. This allows users to navigate between their stored sessions to find information that they have previously stored. The design decisions here were based on findings from our interviews with lawyers. The respondents all said that existing tools did not have sufficient capabilities for tracking and storing information that they found which was relevant to a particular search.

Currently the document editor can support up to sixteen people working on

the same document at one time. Each user is assigned a *write colour* when their account is created on the system. Text that they contribute to a shared document is highlighted in this colour and colours persist with an accessible key. This means that it is easy for other users to see who has been working on a document. Collaboration and document revision is in real time. Users can see what other contributors are doing as they themselves work on a document. The case report window shows the text of the currently selected case or act. This can be *popped out* to form a full screen modal window for ease of reading and navigation. The case report panel has a search facility which highlights matches in the report based upon user keywords. The user can jump forward or backwards between search hits. Citations and passages from the case report pane can be copied directly into the current document by highlighting the text and selecting *Insert into Document* from the contextual menu. The user can search for new cases directly from *Research View* using the links in the top menu, without losing their place in the current document or the chatroom associated with it.

5.13 Key evaluations

In this section, the performance of LARC will be compared in several key aspects to *JustCite*. This is a manually-curated legal research platform which provides similar functionality to LARC. The evaluation here focuses on whether the automated content processing and visualisation systems which are described in this chapter perform competitively with the manual environment. The evaluation will cover three key metrics. In subsection 5.13.2, the case reference identification algorithms are evaluated to establish whether they provide a similar breadth of coverage to the manual system. In subsection 5.13.3, the identification of case treatment is compared between the two platforms. In subsection 5.13.4, the quality of identified interesting phrases in LARC is considered in comparison to the editorial selections of keywords in the manual system. The analysis uses landmark legal cases which are present in both the LARC and *JustCite* databases. These cases have been selected at random from the list at [189].

5.13.1 Differences between LARC and JustCite

As discussed previously, the significant architectural difference between the JustCite legal information system and the LARC prototype that is presented in this thesis is that JustCite relies upon manual curation, editing and asset management

by a team of legally-qualified editors in order to deliver its user experience [46]. LARC is entirely automated and uses unsupervised algorithms for information management, link identification between sources and linguistic analysis functions. This approach has been taken on the basis that an open source platform will not easily replicate the levels of financial investment and expertise gathering that are required in order to reproduce a comparable broadly-qualified user and developer community. The forthcoming sections of this thesis focus on the levels of accuracy and equality of outcome which LARC manages to achieve automatically when compared to the manually-curated JustCite environment.

There are a number of operational and design differences between the two systems which are worth highlighting at this point. Figure 5.14 shows that the initial search interface in JustCite is focussed on identifying seed cases through a text search on case title. Although it is possible to narrow the result set and to interrogate it on different parameters after an initial list has been returned, LARC provides more flexibility by exposing multiple facets for case identification from the start of the search process. This reflects a deliberate design decision from the contextual inquiry to enable tighter groups of case and legislation search results to be returned because users found the large quantity of information which is returned in other products after an initial query to be confusing and to lead to information overload.

Figure 5.15 shows the view of a case report that is returned in JustCite once a case of interest has been selected by the user. This view is notable for the fact that the case report is more effectively formatted and provides more consistent data for parameters such as judge names than is currently the case in LARC. This difference stems from the fact that JustCite sources are cleaned, formatted and edited by teams of legally-qualified experts. However, it is also evident that the level of integration achieved in LARC is superior to JustCite. The manual system does not include features such as a query history; a shopping basket to store interesting information, case and legislation links; or the same level of in-place searching within the text of a particular case report that LARC provides. The ultimate view of information in JustCite is the reading pane for a case report - indeed, legislation is not available in this platform - but LARC uses an integrated document editor to allow users to compose their own work products which incorporate interesting information from the various sources available. Finally, it is also worth saying that the search interface and the reading view in JustCite are segregated whereas they are integrated in LARC. LARC makes it possible to search for new information

directly from Research View, updating only those parts of the interface that need to change when a new case is selected. In JustCite, the user has to come out of the reading view altogether in order to start a new search.

Figure 5.16 shows the case citation view in JustCite. This contains similar information to the augmented citation display in LARC - cases which cite the current case positively, cases which cite the current case negatively and cases where there is a citation which cannot be categorised as either positive or negative. JustCite provides more information than LARC about the jurisdiction of particular cases but this reflects the fact that its database covers material from multiple legal systems whereas LARC presently only operates on English legal data. The difference in design between the two systems here is that LARC implements this information directly in the citation diagram so that the data is presented to the user when they hover over a particular node. It means that there is less need to switch between different views in the application in order to obtain the same level of detail about linked cases. The citation layout in LARC is also more flexible because the data about citations can be obtained simply by hovering over any case at any level in the citation hierarchy. The same level of detail can only be achieved in JustCite after multiple interrogations of the database and consequent switches of context.

The citation layout in JustCite, which is shown in Figure 5.18, was the starting point for the similar visualisation in LARC. The design differs between the two systems quite substantially, however. The figure shows that the JustCite diagram becomes difficult to understand when there are a large number of linked cases to display. The use of space here can be said to be inefficient. LARC overcomes this problem by using a scrolling layout system and by using on-demand node loading so that information is displayed for a linked case only when the user asks for it. LARC also tries to be more flexible by removing the limitation in JustCite that a precedent diagram for a particular case is the ultimate view of the current dataset. In LARC, it is possible to click on any linked case in order to centre the citation layout on the selected source. This means that the precedent layout is responsive to user requirements because the user can change their view of the data as their search requests become more specific and granular. The sentiment colouring in LARC is also different. It assigns a range of colours to graph edges in response to the degree of positivity or negativity in a citation rather than just the red, green and amber classification in JustCite.

Finally, Figure 5.17 shows the interesting or key passages view in JustCite. This functionality is partially replicated in LARC through the Interesting Phrases information which is available for any case in the citation layout. The interesting phrases are not passages here but collocations which appear often in a particular case report, ordered by decreasing frequency. The difference in approach is largely because teams of manual editors in JustCite can effectively annotate key passages whereas LARC relies on automated algorithms for the establishment of linguistic frequency and saliency. Nevertheless, it is argued that the LARC approach is more powerful because the entire Interesting Phrases functionality is linked with a single click into language search mode. This facility is almost completely absent from JustCite. It is possible to search case report text on keywords in the latter system once a case of interest has been selected. However, the language search feature in LARC provides much more flexibility and functionality by exposing the whole text database as a search facet, using measures of saliency to identify cases and legislation of interest by linguistic content from the start of the search process.

5.13.2 Citation Accuracy

The analysis in Table 5.2 shows that LARC performs on a par with JustCite in identifying citations from case reports. There are some instances where the automatic algorithms isolate a greater number of citations than the manual system. This is often caused by the unpredictability of the underlying report text. For example, *Woolmington v DPP* contains several instances where a case citation is given multiple times using different naming conventions. LARC attempts to conflate very similar citation strings using a Levenshtein distance measure. However, cases where a citation is identified in full followed by subsequent single-word references are sometimes not ranked as similar enough to constitute the same citation. Instances where LARC identifies fewer citations in a case are usually caused by the fact that the BAILII sources only contain the judgement of a case.

5.13.3 Treatment Accuracy

In Table 5.3, the analysis focuses on the instances where the sentiment algorithms in LARC rate a case as positively treated in comparison to those that are categorised as “*applied*” under the *JustCite* vocabulary. There are several cases where *JustCite* editors have not processed a case for this information (no treatment data is available). This in itself demonstrates a benefit of the LARC approach. There

Case	JustCite	LARC
Bolton v Stone	4	6
Hedley Byrne v Heller	6	8
Securicor Transport Ltd v Photo Production Ltd	4	2
Woolmington v DPP	8	4
Home Office v Dorset Yacht Co	3	12
Furniss v Dawson	3	3
Associated Provincial Picture Houses Ltd v Wednesbury Corporation	5	4
Equitable Life Assurance Society v Hyman	3	8
R v Kansal	16	34
Fairchild v Glenhaven Funeral Services Ltd and others	51	65

Table 5.2: Comparison of citation coverage in ten key cases: JustCite and LARC

Case	JustCite	LARC
Bolton v Stone	4	4
Hedley Byrne v Heller	4	5
Securicor Transport Ltd v Photo Production Ltd	2	2
Woolmington v DPP	0	2
Home Office v Dorset Yacht Co	2	12
Furniss v Dawson	3	3
Associated Provincial Picture Houses Ltd v Wednesbury Corporation	2	2
Equitable Life Assurance Society v Hyman	0	3
R v Kansal	0	6
Fairchild v Glenhaven Funeral Services Ltd and others	1	34

Table 5.3: Comparison of treatment accuracy in ten key cases: JustCite and LARC

is not always resource or time for editors to deal with cases in a manually-curated system, whereas an automatic algorithm ensures equality of treatment across the available content. The results show a high degree of compatibility between the two systems. LARC is more likely to rate a case positively than *JustCite*. This perhaps highlights the focus here on low-level linguistic features (n-grams) compared to the application of higher-level knowledge in the manual environment. It is also worth saying that LARC is not entirely interchangeable with the *JustCite* treatment system. The gradient-based colouring that is used in the citation diagram offers a more granular appreciation of language about citations.

Case	JustCite	LARC	Matches
Bolton v Stone	5	4	2
Hedley Byrne v Heller	7	7	6
Securicor Transport Ltd v Photo Production Ltd	0	7	0
Woolmington v DPP	6	4	2
Home Office v Dorset Yacht Co	0	7	0
Furniss v Dawson	5	7	5
Associated Provincial Picture Houses Ltd v Wednesbury Corporation	6	10	6
Equitable Life Assurance Society v Hyman	6	7	4
R v Kansal	5	15	3
Fairchild v Glenhaven Funeral Services Ltd and others	4	15	4

Table 5.4: Comparison of keyword category labels in ten key cases: JustCite and LARC

5.13.4 Keyword Accuracy

In Table 5.4, the summary keywords that are assigned by editors on *JustCite* are compared to the interesting phrases that are identified by LARC. The table shows the numbers of keywords assigned by each system and the number of times that a keyword matches between the two. Again, there is a reasonable level of commonality between the two platforms. In general, lower level categories are matched better (such as *asbestos* and *exposure* in *R v Kansal*) whilst higher-level categories are less likely to be evident directly from the text. This is because the LARC system uses a frequency filter to assign the most dominant noun phrases as interesting phrases in a case. A category such as *insurance* might reflect a word that does not appear often in the text as it is a manually-applied summary word.

5.13.5 Differences in analysis between JustCite and LARC: Fairchild v Glenhaven

In the previous section, the entries in Table 5.2, Table 5.3 and Table 5.4 for the case **Fairchild v Glenhaven Funeral Services Ltd and others** are highlighted. This case uniformly features large discrepancies in the citation, treatment accuracy and keyword category label metrics between JustCite and LARC. The difference in citation coverage (**Table 5.2**) arises because case links in JustCite are identified manually whereas the algorithm in LARC is unsupervised. Manual parsing of case text is more accurate than the automated system. The regular expression system

used in LARC encounters difficulty in conflating repetitions of the same citation where different instances are truncated differently. So, for example, whereas one link to *Donoghue v Stevenson* and another to *Donoghue* should not result in two separate links being identified, the two references are not close enough textually for the algorithm to identify both as a reference to the same case.

The same case in **Table 5.3** is rated as positively referenced more through LARC than it is in JustCite. This arises partly because the sentiment system in LARC is more granular than in the manual system. The result shows that the language around 34 citations for *Fairchild* in LARC contains essentially positive collocations whereas the JustCite editors have decided that that same case is *applied* in only one other judgment. There is generally a good correlation between positive collocations around a case citation and instances where it has been directly applied by the presiding judge, but the two approaches are not always directly comparable.

Table 5.4 shows that more important keywords are identified in *Fairchild* through the processing model in LARC than in the manual editorial process for JustCite. This probably arises because the human concept of importance is more selective and less dependent upon statistical measures of frequency than the automated algorithm. It could be argued in this situation that LARC provides a more reliable metric because it operates purely on the number of times a keyword occurs in the text of a judgment. Whatever the interpretation, however, the metric of significance here is that all the manually-identified keywords from JustCite for that case also occur in the result set from LARC.

5.14 Discussion

5.14.1 Mapping design decisions to barriers

The LARC platform has been designed and developed specifically to address the most severe barriers to effective working practice which are created or not addressed by existing legal information systems. See Appendix A, Section A.8 for more details. The issue that **pen and paper notes tend to exist in silos** is addressed by providing a collaborative document authoring and notetaking environment which can be accessed by groups of lawyers at the same time.

The issue that **context switches inhibit computer use** is addressed by implementing an integrated environment for legal research which incorporates document

authoring at its core. Searches for legal information take place in the same context as notetaking and content visualisation.

Information overload is addressed through the application of automatic content parsing algorithms which seek to present high-quality information to the user. This approach encompasses source preparation for functions like faceted search and other features like the identification of important phrases from cases in an easily-navigable visualisation.

The issue that **relevance and precedent hierarchies are not shown effectively** is addressed through the implementation of a citation index for each case and act in the LARC database. The sentiment analysis of cases in citation view also attempts to show visually how important different cases are in relation to the root case of interest.

Current legal information systems make it **difficult to bring notes together**. This problem is addressed in LARC by providing a comprehensive session and source management facility which is implemented as a shopping basket. Session records persist in the system which means that users can navigate forwards and backwards through previous shopping baskets to recall their focus and research direction over time.

The contextual inquiry in Chapter 4 highlighted that **synchronous collaboration seemed to be difficult and sub-optimal**. The document editor in LARC allows for up to 16 people to contribute to a particular work product at the same time. The instant messaging facility enables real-time communication between these contributors. Case links, text passages and other linguistic information can be inserted into the chat with two mouse clicks. This replicates and updates previous research which highlighted the importance of shared search histories in collaborative work.

The three data sources in Chapter 4 which inform the design of LARC all demonstrated that **email is a dominant enabler of group working in the legal domain, but it is not ideal**. This work seeks to reduce the dependence of lawyers on email by enabling real time collaboration focussed around shared documents. The chat facility also presents another more immediate platform for instant communication between members of the legal team.

The solicitors in the interviews in Chapter 4 all complained that it was difficult

to **audit the progress and research directions of junior colleagues** who are most often tasked with legal research work. LARC attempts to contribute here by implementing a timeline visualisation. Senior lawyers can use the timeline to see how work products develop and where research work is heading at any point in time. The timeline is augmented with saved contexts which replicate the state of research view exactly at important waypoints.

There was a sentiment amongst students in the contextual inquiry that **general purpose research tools were often better** - easier to use and more accessible - than specialist information platforms for lawyers. LARC attempts to address this in two ways. Firstly, annoying user interface problems with other products are tackled by the implementation of a customisable and switchable design in LARC. Secondly, LARC integrates content from an accepted general purpose legislation database which is provided by the UK government.

5.14.2 Limitations and opportunities

A significant limitation to the uptake of the LARC platform is the size of the case law database. This is introduced by the limited availability of open legal data. The BAILII resources that have been licensed for this research comprise only 70,000 case reports. Coverage limitations introduce a need to switch context away from the platform which is one of the key barriers that LARC tries to overcome.

Lawyers have come to expect a highly-processed form of legal information. This means the provision of head notes, paragraph referencing and cross-referencing, abstracts, page tracking against printed versions and other editorial content. The argument in this thesis that open access legal data tied to open source software tools can help to diversify the sector rests on the idea that expensive manual intervention can be replicated in many aspects by automated machine processing. This works to a large extent but cannot entirely replicate the expected annotations.

The Ruby On Rails framework which LARC runs on is a good choice for fast development but it does not scale particularly well. Ruby is an interpreted language that is approximately 50 times slower than C code [18].

In order to make a difference in legal education and training, LARC needs to be tested and adopted in live pedagogical environments. This would be a gradual process, helped by the fact that LARC is an open access platform with a low barrier to entry. Legal curricula are changing now to adopt next generation tools such

as *Casetext* and *Ravel Law*. There is scope for the adoption of another new tool, particularly one which has been informed and designed so specifically for mooted activities. The other part of this issue is that the system needs to be adopted and developed by computer scientists and software authors who are interested at least to some extent in the law. The “big data” nature of the domain should ensure that involvement will appeal to different types of contributor in the open source community.

5.15 Conclusion

In this chapter, the LARC platform for integrated legal research and digital notetaking in groups has been presented and described. This software and its key design decisions form **Contribution C1.2** of this thesis, as enumerated in Chapter 1. Nine key barriers to effective working practice that current tools in this sector introduce have been carried forward from the three surveys in Chapter 4. A process of iterative paper prototyping was undertaken before starting to develop the system. This was done in order to ensure that design decisions taken in the software directly address the barriers that have been found. LARC has been designed to address the most severe barriers through a focus on collaboration; reducing the problem of information overload; enabling audit of work by tutors and senior lawyers and creating a curated research environment by preserving the idea of precedent and case hierarchy.

An initial evaluation shows that the platform performs well in terms of coverage and parity of presentation with a manually-curated tool. Priorities for future work revolve around an emphasis on preserving the open access to legal data provided through LARC and the open source status of the software platform. Challenges include content completeness and building a development community around the platform. Future work on the platform will enhance the simplicity of the interface and augment the visualisations available to better support legal work.

III

PART III

EXPLORING LEGAL LANGUAGE

INTEGRATING LANGUAGE SEARCH

6.1 Thesis process

This chapter contributes to answering the main research question in this thesis by addressing the following step of the process outlined in Table 1.2:

- **P5 - Consider how search interfaces for linguistic content can be designed to target end-users who are not linguists or computer programmers.**

This step in the process will be addressed through a description of the algorithms from corpus linguistics which have been implemented in the **LegAl Research and Collaboration** platform. The software includes dedicated facilities to run language queries against large corpora of case law reports and legislation. The manner in which this integration with the legal research functionality that was described in Chapter 5 is achieved results, for the first time, in a detailed set of guidelines for the use of corpus-based tools in platforms which are not primarily aimed at linguists.

6.2 Introduction

In this chapter, the integration of language search functionality into LARC which runs on corpora of case law and legislation will be introduced, discussed and justified. Initially, the need for new legal corpora will be examined. This is justified

in terms of the limited size of existing resources and a lack of focus currently on corpora which feature sources from English law. Fundamental guidelines for the creation of large-scale legal corpora are proposed with reference to particular standards that have been established by scholars and practitioners in corpus linguistics.

The search for a fundamental unit of linguistic meaning is considered next. The approach here is based upon the idea that words do not have atomic meanings. Meaning is derived from common contexts of usage. The system of legal precedent operates in a similarly contextualised manner and therefore lends itself to evaluation through corpus-based techniques. This is important because it informs the way in which information should be extracted from corpora and presented to the user in an iterative search interface. Different algorithms for identifying and ranking units of meaning are presented.

The language search interface in LARC implements an iterative search methodology which focuses on drilling down through collocation data until a relevant linguistic construction is discovered and isolated. The user then moves from examining collocations to a concordance view with larger contexts around the identified search node. The utility of the concordance as a search interface is discussed. Various algorithms for organising, sorting and viewing concordances are proposed. These are novel to the LARC platform in their implementation and extension. The limitations of a traditional Keyword in Context concordance model are highlighted and new processing steps for elucidating concordance data and for visualising that data are introduced. Traditional systems for the analysis of corpora bring with them several methodologies which, when applied to software for non-expert users, are problematic. These problems are identified and discussed. LARC implements appropriate mitigations for the issues that are identified.

The screenshot displays the RAVEL Law search interface. At the top, a search bar contains the word "negligence". To the right of the search bar, a navigation menu includes links for "Prints", "Cites", "Judges", and "Statistics". Below the search bar, a timeline visualization shows the distribution of search results from 1793 to 2008, with a peak around 1980. The main content area lists five search results, each with a rank number, case name, citation, and a brief description of the case's relevance to the search term "negligence".

RAVEL negligence All Federal and State

843,087 Matching Results

Visual Search List View

1793 2008

1. **Estelle v. Gamble**
425 U.S. 971 U.S. Supreme Court | November 30th, 1976
Cited by 30,691 opinions

MATCHES

- ... the doctors may be guilty of nothing more than **negligence** or malpractice. On the other hand, it is surely ...
- ... In Andersonville were the product of design, **negligence**, or mere poverty, they were cruel and inhuman ...

2. **Farmer v. Brennan**
511 U.S. 825 U.S. Supreme Court | June 6th, 1994
Cited by 28,878 opinions

MATCHES

- ... describes a state of mind more blameworthy than **negligence**. In considering the inmate's claim in *Estelle* ...
- ... indifference entails something more than mere **negligence**, the cases are also clear that it is satisfied ...
- ... indifference lying somewhere between the poles of **negligence** at one end and purpose or knowledge at the other ...

3. **Ashcroft v. Iqbal**
556 U.S. 662 U.S. Supreme Court | May 18th, 2009
Cited by 17,903 opinions

MATCHES

- ... position wrongs, or for the nonfeasances, or **negligences**, or omissions of duty, of the subagents or servants ...

4. **Beil Atl. Corp. v. Twombly**
550 U.S. 544 U.S. Supreme Court | May 21st, 2007
Cited by 146,097 opinions

MATCHES

- ... IN FOOTNOTES ...
- ... contrasts sharply with the model form for pleading **negligence**, Form 9, which the dissent says exemplifies the ...

5. **Celotex Corp. v. Catrett**
477 U.S. 317 U.S. Supreme Court | June 25th, 1986
Cited by 140,635 opinions

Page 1 of 117

Figure 6.1: The standard method of language search presentation in Ravel Law.

6. INTEGRATING LANGUAGE SEARCH

The screenshot displays a web-based legal corpus search interface. At the top, a navigation bar includes links for 'Search Cases', 'Search Legislation', 'Language Search' (highlighted), 'My Collection', 'Document Selection', and a user profile 'Evan Brown'. Below the navigation bar, the main search area is divided into several sections. On the left, there are input fields for 'Search collection:' (with buttons for 'Case Law' and 'Legislation'), 'Query history:' (with a 'Clear History' button), and 'Key Sources'. To the right of these fields, a status bar indicates 'Found 200 open collocates - 0 closed collocates' and features a 'Clear query' button. Below this, a toggle switch for 'Collocates' is set to 'On', and a 'Concordance' button is visible. The main results area is split into two panels. The left panel, titled 'Open Collocates', lists various words associated with 'negligence', such as 'contributory negligence', 'negligence unskilfulness', 'negligence fellow-serv...', 'proferens negligence', 'negligence discounten...', 'indemnitee negligence', 'negligence indolence', 'negligence engine-driver', 'negligence misconducts', and 'negligence incrementally'. The right panel, titled 'Closed Collocates', states 'No significant closed collocates found.' Both panels include pagination controls at the bottom, showing 'Prev', '1', '2', '3', '4', '5', '6', '7', '...', '19', '20', and 'Next'.

Figure 6.2: The corpus-based approach to presenting language search results.

Search Cases ▾ Search Legislation ▾ Language Search

70580 cases - 73510 statutes | My Collection ▾ Document Selection ▾ Evan Brown ▾

Query history: Clear History

Key Sources

Show 10 ▾ entries

Search:

Case Title	Case Citation	Result Frequency	Citation Index
Investors Compensation Scheme v. West Bromwich Building Society	[1997] UKHL 28	3	868
English v Emery Reimbold & Strick Ltd.	[2002] EWCA Civ 605	1	434
Johnson v. Gore Wood & Co.	[2000] UKHL 65	23	404
Allied Maples Group Ltd v Simmons & Simmons (a firm)	[1995] EWCA Civ 17	9	383
H & Ors (minors), Re	[1995] UKHL 16	1	381
Swain v Hillman & Anor	[1999] EWCA Civ 3053	1	360
Assicurazioni Generali SpA v Arab Insurance Group (B.S.C.)	[2002] EWCA Civ 1642	6	343
Pozzoli Spa v BDMO SA & Anor	[2007] EWCA Civ 588	1	305
Caparo Industries Plc v Dickman	[1990] UKHL 2	22	301
Three Rivers District Council v. Governor and Company of The Bank of England	[2001] UKHL 16	20	299

Showing 1 to 10 of 2,250 entries

Previous 1 2 3 4 5 ... 225 Next

Figure 6.3: The key sources dialogue which is extracted from the corpus in LARC.

Search Cases ▾

Search Legislation ▾

Language Search

70580 cases - 73510 statutes |

My Collection ▾

Document Selection ▾

Evan Brown ▾

Key Sources

Found 41312 concordance lines

Collocates

Concordance

Clear query

Sort: None ▾ on token: 1 ▾ for context: Left ▾

Sort Concordance

Explore Concordance

Export Concordance

☒ Normal rank

☐ Typical rank

☐ Typical with filter rank

■

express right is granted to the owners of No.5 over the strip . Absent gross

negligence

■

the transfer , informed Mr and Mrs Mainzs solicitor but the latter , through gross

negligence

■

, may be told that he has lost his right by his delay and his

negligence

■

not make this alteration without the proprietors consent unless he has either by fraud or

negligence

■

, may be told that he has lost his right by his delay and his

negligence

■

, may be told that he has lost his right by his delay and his

negligence

■

visited with personal loss on account of mere errors in judgment which fall short of

negligence

■

defendant was negligent . It is also common ground that the damage resulting from that

negligence

■

1995 . The defendant says that it was prior to 1st September 1995 . The

negligence

■

the claimants now sue , and which the defendants concede was caused by their actionable

negligence

■

. Neither the claimants nor Provident Mutual became conscious of this fact until 1995 .

negligence

Prev 1 2 3 4 5 6 7 ... 4131 4132 4133 Next

Ad-hoc Classes

Figure 6.4: The concordance view for a query on *negligence* in LARC.

192

6.3 Designing a corpus for use by lawyers

6.3.1 Why use a corpus linguistics approach?

Figure 6.1 shows the initial results screen for a query on *negligence* from Ravel Law, which is one of the major existing legal information platforms. By default, search results are presented in a list. The layout displays cases that the system deems to be relevant to the query apparently in descending order of importance. However, the metric used to calculate how the individual search hits are scored for relevancy is unclear. The second case in the list, for example, has more hits in the text for negligence than the case that is ranked first. The available context around search hits is small and apparently static so that it is difficult for the user to understand and to synthesise information about why a particular case is deemed to be relevant without reading the whole case report.

There are also few options to filter and organise search hits after the initial query has been made. A general query on *negligence* can in fact be focused much more specifically on detailed areas of legal treatment, such as judicial reasoning around *contributory negligence* or *negligence* as it relates to the doctrine of *volenti non fit injuria*, for example. If the user clicks on a result, they can read the full text of that particular report. This view is well formatted with paragraph references, cross-references, notes and highlights on important text, but it does not by default highlight areas of the text that are relevant to the query which first led to discovery of the case.

As a comparison, Figure 6.2 shows the initial response to a query for *negligence* through LARC. The software displays common collocations for negligence, or words which appear alongside *negligence* more often than chance itself would dictate, ordered simply by the frequency with which the collocation occurs. This display takes into account different windows, or gaps, between the node word and the collocation. Although the Ravel Law software makes it possible to query that database for *negligence* occurring with other words in the same hit environment, this relies on a boolean search language and is unguided. It is suggested that the guided approach taken in LARC, which means that common collocations for *negligence* are automatically identified from the corpus and presented to the user by frequency, is more accessible. The practice of guiding the user by automatically suggesting additional words for their particular query which are based on the textual environment of *negligence* should also guarantee a higher degree of search

success and overall relevance. The user can see that *contributory negligence* is an important legal topic associated with their query. They can also drill down in an unlimited number of steps to focus the query on *contributory negligence* and then, progressively, on *contributory negligence* as it relates to the offence of *battery*, for example.

In LARC, important cases which feature the term *negligence* are immediately displayed in a collapsible panel once the query has been made. The metrics by which these cases have been ranked for relevance are transparently shown in the table and the results can be ordered either by how often a case has been cited or by how many hits for *negligence* occur in the report. The table can also be searched in real time through the text field in the header. The user may also switch to a concordance view at any time, which gives the context of the word *negligence* in different cases. These contexts are dynamic and can be lengthened by requesting a full co-text for any particular result. In this way, the user can ascertain the relevance of a particular case or item of legislation without having to read the full text of the case report.

6.3.2 The need for new legal corpora

As has been discussed in Chapter 3, there is an increasing move to publish legal information for open and public access. See Chapter 3, Section 3.2 for a more detailed discussion of this development. For example, The Case Law Access Project has digitised and provided access to three hundred and sixty years of United States case reports, covering some 6.4 million individual cases. This is the product of a joint initiative between Harvard Law School and Ravel Law [26]. The available database can be queried through an API and some elements of the collection can be downloaded as XML or plain text.

In this context, it is legitimate to question whether additional corpora of legal data are required. However, most of the available open access collections of case law cover only the United States. The situation for English law is much more constrained. Existing collections like the Old Bailey corpus or the British Law Report corpus are either too small and limited in time to be representative of the domain or they are composed of materials which are entirely historic. The data licensing terms for initiatives like The Case Law Access project are also restricted to non-commercial use.

At present, the database is not available for download in either XML or plain text as a whole. The project offers ZIP files of the content for Illinois and Arkansas but usage of the majority of the data is subject to bespoke licensing terms. An application like LARC could not operate simply within the confines of a pre-determined API. The platform requires access to archives of plain text case law and legislation in order to derive processed data from the raw materials. Thus there is a need for true open source development in parallel with a clear requirement to improve open access to and coverage of the products of English law. This should mean that available data encourages diversification in research and product development in the United Kingdom in a manner similar to that which is starting in the United States.

One of the key aims in designing and developing language search in the LARC platform was that it should be based upon a living corpus of linguistic data. A living corpus is a body of text which is not unnecessarily limited in scope by time. This means that both historic and contemporary sources are treated and considered equally in the composition of the corpus. It also means that the various preparation tools which are required for corpus creation be capable of updating the resource as new data becomes available. This ability to revise corpora should be achieved in a manner which is technically sustainable, such that updates can be made without unnecessary processing overheads which may impact upon update frequency and granularity.

The LARC corpus includes case law reports from an archive maintained by the British and Irish Legal Information Institute (BAILII) and legislation from The National Archives (TNA). The total size of the available text for language search in case law is 423,335,518 words at the present time. The total size of the legislative corpus is 210,181,611 words. This makes the LARC database the largest formally-constructed corpus of English legal sources by some distance at the moment. This means that the dataset is the biggest collection of legal materials which has been processed, augmented and then codified for use by corpus management software. In terms of outright size, the LARC database is smaller than the archives of both Lexis and the Incorporated Council of Law Reporting, but this material could not be used for corpus interrogation without significant extra work. Part of the novelty in the LARC system is that it represents the first time that corpus creation and management techniques have been used on such large collections of information from the English legal domain. Many of the design decisions that were taken

in developing the platform have been implemented because of the size of these materials and their consequent demand for storage, processing and interrogation resources.

6.3.3 The argument for clean text

A key design decision in the creation of corpora for use in LARC was that the resources should be plain text. This aligns with a theory from corpus linguistics that sources should be as unadorned with extrapolated data streams as possible. It means that the results of any processing of the corpus should not be codified in the corpus itself unless such codification is unavoidable. The justification for this approach is that many initiatives to tag, segment, describe or otherwise elaborate plain text are specific to a particular interrogation requirement. The presence of such parallel data streams in the corpus file at best complicates the underlying material unnecessarily and, at worst, renders the corpus unusable for future applications without a lengthy and expensive re-codification step. This feeds into a larger move away from predefined data classifications and imposed contexts. An ontology that could be created on legal data for use by lawyers is unlikely to be useful or relevant to a linguist who is interested in legal language, for example.

The LARC corpora for both case law and legislation were built from a vertical stream of the individual source tokens (words) themselves. This was augmented with a second stream which contained Part of Speech data for each token. More information about the preparation process for the corpora is provided in Section 6.3.5 of this chapter. This meant that each corpus was plain text and that the sources were not segmented or classified in any way. The use of one monolithic corpus instead of multiple smaller corpora for different years or courts, for example, was preferred. This design decision keeps user interfaces for interrogation simpler because there is no need to provide facilities for switching between sub-corpora or for looking at the search results from different sections in parallel.

6.3.4 The problem of structure

There is a tension in LARC between the systematised, structured presentation of information for facilities such as faceted search and the unstructured, unadorned data that is available from the corpora. The different data stores for LARC exist in isolation. This means that there is a MySQL database of case law and legislation

which is organised for discovery by instrument name, date, court and so on. There is also a separate binary store for the unstructured corpora. Early in the prototyping stage, it was found that this created problems in implementing certain important functionality, like being able to show the user which case or statute a particular linguistic hit comes from. The rationale for the platform is to provide an integrated system for legal research and, as such, it is important to be able to tie language returned from the corpus back to case names and statute details. This binding ensures that there are no dead ends in the system. A user can select the context of a specific hit for the query *negligence*, for example, and they can see which cases that language appears in. They can also select any of the cases to view the complete case report in *Research View*.

Integrating the LARC platform in the manner described above required a compromise between the desire that corpora should be essentially plain text and the need to include some structure in the store so that language can always be tied back to the sources that it comes from. Initially, the corpora were segmented by document and a header was included before each individual case report which provided the canonical `case_page_id` or `legislation_page_id` identifier from the database. It was then possible to query the corpus for language, to find the correct case identifier for each result and to furnish essential information like case name, citation and date from the MySQL database.

However, as the corpus grew and the information that was being extracted from it became more complex, it transpired that repeated database queries of the faceted information upon every corpus interrogation were unsustainable. Query times across the 400 million words of case law for common words like *negligence* took several minutes to return results, for instance, because the MySQL store was being accessed several times for every hit. As a result, a decision was taken to include the following information in segmentation headers directly in the corpus store itself: `case_title`, `case_citation`, `case_page_id`, `case_court`, `case_judges`, `case_date`, `star_ranking` and `citation_index`. The encapsulation of this information directly in the corpus allowed for it to be extracted using queries on the binary indices of the unstructured store, which led to a query for *negligence* completing in several seconds rather than several minutes.

Another problem related to structure in corpora is that they must have an interface for interrogation. This necessity in itself means that several query languages have been proposed which allow users to specify the information that they are interested

in retrieving. There is no absolute standardisation here but most systems employ a query language which is similar to Corpus Query Language (CQL) from the SketchEngine product [157]. CQL is a structured language for formulating search requests against corpora. It standardises the format of queries and defines a finite vocabulary for searching against different properties that have been compiled in the corpus files. The central advantage of CQL is that it has become widely used and accepted. It has replaced the older environment where every corpus manager implemented its own query language almost entirely.

As a practical matter, a design decision was taken to use the open source Manatee corpus manager which is a customised part of SketchEngine itself. Manatee was made open source during the PhraseBox project under the guidance of John Sinclair. It is essentially a freely-available version of the corpus manager in the commercial SketchEngine product but it has some limitations when compared to the full version. None of these limitations were significant for the LARC project, however.

`[word="contributory"][]{0,1}[word="negligence"]`

Figure 6.5: A simple CQL query for *contributory negligence*.

The complexity of Corpus Query Language reflects the fact that it is designed to be used by linguists and computer programmers. Even the relatively simple query for *contributory negligence* that has been considered previously must be encoded in a specialist manner in CQL before the corpus manager will return results, as can be seen in Figure 6.5. More complex queries for lemmas (word roots), specific Part of Speech tags and wildcards require an understanding of regular expressions. The LARC user interface hides this structure and complexity entirely. It provides simple controls and parameter menus from which a conversion to an appropriate CQL query is achieved transparently.

6.3.5 Preparing the LARC corpora

In order to create corpora for use in the Manatee corpus manager, the raw text of case reports and statutes has to be formatted and encoded correctly. This involves creating a single file for all the text in the entire reports database. The program code for preparing the two corpora can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>. The code first

creates an empty file for the plaintext corpus. It then iterates over every case record in the `case_pages` table and retrieves the corresponding entry for case data. The simple HTML record of an individual case report is sanitised to remove line breaks and other unwanted content.

Next, the textual report content is written to a temporary file on disk. Once that has been done, the case title, primary citation, court, case date, case judges and citation index data are retrieved. If there are multiple judges for a particular case, each name is taken and concatenated together with comma delimiters in a string. All of this data is then written to an XML section definition line which follows a format for custom data attributes that is laid down by the Manatee software. For each case, the structure delimiter is written to the main file first. The plain text of the associated case report is then fed through a Part Of Speech tagger. This step accomplishes two things.

First, it chunks the text into individual tokens. Each token is then placed on a single line of the output file. This means that the case report becomes a long vertical collection of single words. The second part of the process inserts a tab character after the end of each word line. A Part of Speech tag for that individual word is then appended. Finally, a structure terminator for the individual document is placed at the end of the file. This file is then added to the end of the main vertical corpus file. Once all the case reports have been processed, the result is a large text file with a column of words and a column of speech tags, interspersed with document separator lines in XML format.

```
<doc case_title="# (A Child)" case_citation="[2010] EWMC 75" case_page_id="1"
case_court="Magistrates' Court (Family)" case_judges="none" case_date="2010-01-01"
star_ranking="0.0" citation_index="0">
```

This	DT	this
decision	NN	decision
is	VBZ	be
part	NN	part
of	IN	of
the	DT	the
Family	NP	Family
Courts	NPS	Courts
Information	NP	Information

Figure 6.6: Sample vertical text output from the corpus preparation program.

The first few lines of the vertical text file for the case law corpus can be seen in

Figure 6.6. After the file is complete, a configuration script for the Manatee system must be created. This configuration is individual to a corpus. It tells the system where to find the appropriate vertical text file that was created in the previous step. It also describes what each column of data in the file represents and how the system should encode the different streams of information. The vertical text file and the configuration script are then used to create binary files and associated indices so that a searchable corpus is generated.

6.4 Moving away from keywords to collocations

Sinclair posits two different models for explaining how meaning arises from language text [162]. The open choice principle suggests that people write and speak on the basis of making a very large number of complex choices about which words to use. At each point where a word or other grammatical structure is selected, a large range of choice opens up and the only constraint is ensuring that the completed utterance remains grammatical. For many years, this idea of open choice (which is also referred to as the “slot-and-filler” model) formed the accepted way of seeing and describing language. Text consists of a series of slots which have to be filled from a lexis which satisfies local restraints. Word choice at any slot position is virtually unlimited. Almost all accounts of grammar in language operate on the basis of this principle of open choice. The second model for building meaning which Sinclair proposed is called the “idiom principle”. This was based on empirical investigations using corpora which made it clear that words do not occur at random in text. The open choice principle therefore does not provide for suitably substantial constraints on the choice of consecutive words.

The importance of linguistic context in understanding and interpreting the law is well illustrated by the concept of “*obiter dicta*”. These are comments around a judgement routinely made by judges that are not in themselves statements of the operative legal principle in a case, though they may be persuasive in future cases, but which may deal with hypothetical circumstances or which are used for purely illustrative purposes to clarify the meaning of the “*ratio decidendi*” - the judge’s decision of principle in a case. They are valuable guides to the judicial reasoning that led to the final decision and are thus integral to the meaning of that judgement. The corpus linguistics approach recommended in this thesis most effectively elucidates the meaning of such judgements by exposing and illustrating linguistic contexts.

“Wittgenstein’s remarks on the grammatical nature of understandings lead inexorably to the conclusion that conceptual understanding is social...The meanings of words can only be understood if we understand the purpose or ends of the human activities of which words are part...Such a theory would, at a minimum, enrich our institutional ontology beyond a conception of law as “rule” and “principle”. More importantly, it would endeavour to show how the faculty of judgement is always at work in legal reasoning without being reducible to schematic rule or principle.” [146]

The idiom principle holds that things that are conceptually related occur in close proximity to one another. There are sets of overarching linguistic choices which can be seen to condition, and to thereby massively reduce, subsequent linguistic choice. Other language structures such as sets and comparative or contrasting series also serve to organise text along predictable and recurring lines. In practical terms, this means that people speak in idioms. These idioms are most often large blocks of language that are chosen at one time. Once chosen, they necessarily dictate a smaller degree of freedom to complete an utterance than the open choice model would suggest.

Sinclair proposes that the smallest indivisible unit in these idiomatic sequences is the collocation. A collocation is a group of two words which can be seen to occur together more frequently in text than would be dictated by chance itself, or by the idea that words are fitted together in a largely open and unrestrained context. At their simplest, the ideas of idiom and of collocation can be demonstrated through an apparently simultaneous choice of two words, like *of course*. This phrase operates effectively as a single word. We do not consciously choose *of* and then decide subsequently to complete the utterance with *course*. This is a single building block to express meaning which is chosen at a single point. Another slightly more complex example which illustrates the point is the construction *set eyes on*. If words collocate significantly, then to the extent of that significance, their presence is the result of a single choice. In practical terms, our implementation of the idiom principle relies heavily on associated linguistic ideas like metaphor and analogy. It is no accident that everybody has a stock of known constructions that have become accepted as whole utterances - like *it is not rocket science* or *went down like a lead balloon*.

In practical terms, the idea of the idiom principle has implications for the way

that information retrieval systems derive results and then present them to users. If the collocation is held to be the smallest indivisible bearer of meaning, then search strategies which concentrate on highlighting single keyword matches in text are deficient. Indeed, boolean search capabilities are also inappropriate because, although this can help to highlight groups of words which occur in proximity to one another, they do not tend to impart information about the importance or strength of any co-occurrence. A new approach is required which takes the initial query and then offers different environments to the user in which the input word features. These environments should be ranked and displayed in terms of the strength of the word associations which they contain.

As an example, the query for *negligence* which has been considered previously is informative. A keyword-matching approach for this query will identify all sequences of text which contain that word. The results may or may not be organised by the frequency with which the query occurs in different documents. However, there is a broad scope of meaning and ultimately of judicial treatment in the area of negligence. For example, the term encompasses the related areas of *gross negligence*, *comparative negligence*, *contributory negligence* and *vicarious liability*. As it stands, a basic system which matches keywords could not elucidate all of these different contexts for the query clearly. The user would likely be forced to run multiple consecutive requests to find material related to each different aspect of their area of interest. This leads both to large result sets, which were found to be confusing in the studies from Chapter 4, and to a requirement for significant manual work in order to identify relevant material from the background of keyword hits. However, utilising collocations to provide a guided identification process which accurately isolates legal information needs can answer real problems identified in the literature and in previous surveys of lawyers and legal librarians, as the following quote from Mishkin demonstrates:

“Three of the skills perceived as most lacking, were related to the construction of an effective search. 58% of the non-academic law librarians felt that new joiners lacked the ability to select appropriate search terms, and 65% felt they were unable to construct an appropriate search string (by combining terms or using connectors), whilst the greatest weakness identified (by 72% of respondents), was trainees’ inability to select the correct resource to answer a query. Given that these skills are so fundamental to the construction of a legal information search it is concerning that so many librarians found them to be

lacking. Addressing these weaknesses in particular, needs to have a stronger emphasis in training programmes across the sectors.” [134]

6.4.1 Extracting collocations from the corpus

The Manatee corpus manager makes no assumptions and is agnostic about which algorithms are used to calculate the strength of association between two or more words in a corpus. By default, a query like the one for *contributory negligence* in Figure 6.5 will return a concordance of search results based on all instances in the corpus where *contributory* appears within eleven words to the left of *negligence*. It is up to the developer of the implementing software, in this case LARC, to create appropriate algorithms which identify the important collocations of *negligence* that feature *contributory* in the returned concordance. Thus, several algorithms have been applied in the course of this research in order to turn the content of unranked concordances from Manatee into collocation displays that are organised by frequency and, ultimately, the strength of association that the text supports between any two words.

The simplest way of evaluating whether a collocate is relevant or not is to compare its observed frequency of occurrence in the concordance with what would be expected given how common the word is generally. It is important to understand here that predicted frequency is not based upon probability calculations. If the source corpus is indeed well constructed and can be said to be large enough to be representative of language from a given domain, there is no need to estimate frequency metrics. The expected frequency here is thus the total frequency of the word in the corpus as a whole, whilst the observed frequency is the frequency of the word in the concordance returned for a particular query. The output of this algorithm is as shown in Figure 6.7.

$$s_c = \frac{f_{conc}}{\left(f_{corp} \cdot \left(\frac{size_{span}}{size_{corp}} \right) \right)}$$

Figure 6.7: The observed and expected ratio equation for collocate significance in LARC

The simple significance of a particular word in the environment of a node word is computed here by taking the frequency of the word in the concordance that is

returned by Manatee and then dividing it by the frequency of that same word in the corpus as a whole. The frequency in the corpus is multiplied by the size of the environment of each hit in tokens (words) divided by the size of the corpus. The environment size is used as a factor to try to weight scores appropriately according to the relative sizes of the corpus and the context that has been returned for each hit in the concordance. This implementation detail is often called *lexical gravity*.

In theory, this algorithm classifies collocates of a node word with a score of greater than 1.0 as more significant than would be expected, and those with a score of less than 1.0 to be less significant than expected. This simple algorithm is one of the most defensible measures of significance because it relies purely on frequency information from the corpus and the extracted concordance. In practice, however, many words in the corpus are quite rare and do not appear regularly in relation to other words. This means that their expected frequency will be low. However, if the word occurs once in the environment of a given node, the corresponding significance score will be large (1 divided by a small number).

The problem with elevated significance scores from the simple calculation presented in Figure 6.7 means that several other algorithms have been proposed for ranking collocates by their strength of association with a given node (or query) word. One of these alternatives which is implemented in LARC is the *z-score*. The algorithm for calculating the *z-score* for a particular collocate of a given node word subtracts the frequency of the node in the corpus from the frequency of the node in the concordance. The result is divided by the standard deviation of the frequency of occurrence for the word in the whole text. All the steps taken to calculate this significance score are shown in Figure 6.8. The results of using the *z-score* significance metric on a corpus generally mean that there are fewer rare words elevated to a position of significance in the output.

$$p = \frac{(f_{corp})}{(size_{corp})}$$

$$\sigma = \sqrt{size_{corp} \cdot (p \cdot (1 - p))}$$

$$s_c = \frac{(f_{conc} - f_{corp})}{\sigma}$$

Figure 6.8: The z-score algorithm for collocate significance in LARC

Another significance measure which is implemented in LARC is the *t-score*. The *t-score* is similar to the *z-score* but the calculation for sigma is replaced by an approximation. Thus the square root of the observed frequency is used as the determinant in this equation. This essentially gives some of the benefits seen in z-score ranking without the complexity of the underlying algorithm. To generate the *t-score*, the expected frequency of a collocate in the corpus is subtracted from the observed frequency in the concordance. This result is then divided by the square root of the frequency of the collocate in the concordance. The full equation used here can be seen in Figure 6.9. Although the results from t-score ranking are generally more reliable than with the observed and expected ratio, there are more function words (which are highly-frequent but not meaningful collocates) at the top of the resulting list.

$$s_c = \frac{(f_{conc} - f_{corp})}{\sqrt{f_{conc}}}$$

Figure 6.9: The t-score algorithm for collocate significance in LARC

LARC also implements the *Mutual Information* algorithm to score collocates in language search. Mutual information [31] was first proposed as a significance measure in the domain of information theory some seventy years ago. The basis for this methodology is that finding a particular word in a sequence of text gives valuable information about what word will come next. Its selection operates as a local constraint. The information value of a token increases as more information

is gained from the presence of a particular word because its role as a constraint on subsequent choices is more pronounced. For example, the word *the* has a low information score because it is difficult to guess what word will come next. *The* does not operate as a coercive constraint on subsequent language choice. However, the word *rasher* has a high information value because it is likely that *bacon* will follow, probably after the intervening word *of*.

The mutual information algorithm can be seen in Figure 6.10. The formula involves calculating a logarithm to the base 2 because the size of a stream of digital data is measured in bits. In fact, that element of the equation is of little importance for ranking collocates since removing it from the formula maintains the rank order but changes the magnitude of the individual word scores. Mutual information tends to promote rare words in the collocate list, including proper nouns, which is usually undesirable. However, this may highlight developing areas of the law or to higher-level classifications in case reports which are not easily identified by filters that promote high linguistic frequency.

$$s_c = \frac{\log\left(\frac{f_{conc}}{f_{corp}}\right)}{\log(2.0)}$$

Figure 6.10: The mutual information algorithm for collocate significance in LARC

Role	Collocate	Not collocate
Node	A	B
Not node	C	D

Table 6.1: The contingency table for log-likelihood calculations.

The last significance algorithm which is implemented in LARC is *log-likelihood*. This measure of collocation strength was first proposed in [53]. The algorithm was presented as a replacement for existing salience measures which, it was claimed, are based upon poor statistical calculations that are difficult to verify in practice and to defend. Likelihood ratio tests are claimed to yield good and defensible ranking results even with small collections of text to work with. The algorithm

starts by taking a collocate candidate token and computing four values for the word, one for each entry in the contingency matrix in Table 6.1. These figures represent the likelihood of the word being a node and a collocate (the idea of self-collocation), a node but not a collocate, a collocate but not a node and neither a node nor a collocate (which means that the word does not appear in the given concordance at all). These contingency values are then used as the basis for a salience calculation which is given in Figure 6.11, where the terms i and j are the likelihood factors from the two pairs of statistics calculated from the contingency table.

$$\begin{aligned}
 O_i &= f_{conc}(coll) \\
 O_j &= f_{corp}(node) - f_{conc}(coll) \\
 E_i &= f_{corp}(coll) - f_{conc}(coll) \\
 E_j &= size_{corp} - (f_{corp}(node) + f_{corp}(coll) - f_{conc}(coll)) \\
 s_c &= 2 \sum O_{ij} \log \left(\left(\frac{O_{ij}}{E_{ij}} \right) \right)
 \end{aligned}$$

Figure 6.11: The log-likelihood algorithm for collocate significance in LARC

All of the algorithms for ranking collocates which are implemented in LARC suffer to a greater or lesser extent from elevating very common words which are not especially informative. This applies particularly to a category of words which can be said to belong to a *closed class*. There is a loose translation here between words which actively invite the collocation of other words to bear meaning and valuable information (known as *open class* words) and those which do not readily impart meaning even with the addition of other words to their immediate environment (*closed class* words). Words in the *closed class* include *if*, *but*, *the*, *a* and so on. It is possible to appreciate that the addition of first-level collocates to these words will convey little relevant information.

The language search interface in LARC segregates *open class* and *closed class* words into two different panels of the screen layout. This is done to ensure that no data is hidden from the user, but that highly-frequent grammatical words which are not informative by themselves do not pollute the significance rankings. The design decision here is to allow each entry on the collocation list for a query to impart as much useful information as possible so that the search environment can be guided and information overload is kept to a minimum.

A static list of *closed class* words was compiled which totals about 400 tokens. The collocate lists are then scanned for these words and, whenever any of them are found, they are placed in the closed list. Those collocations with high information content are thus prioritised in the language search interface, as shown in Figure 6.2. It was decided that the placement of a collocate to the right or left of the node word would be shown to the user in language search. Most corpus systems operate on word lists which do not denote relative positions. This is confusing to users who are not linguists and it requires an additional mental step to process the result.

6.4.2 Search as a workflow - drill down

An important design decision in the LARC interface is that search must operate as an iterative workflow. This means that the user is provided with capability to progressively change and focus their search requests as they find more information from the corpus. Thus, a query for *negligence* returns a list of *open class* words with the collocate *contributory* ranked as the most statistically significant token to the left of the node. The user can then click on the context menu to the side of that search result and select an option to *Drill down*. LARC then performs a search for *contributory* as it appears to the left of *negligence* in the corpus. This functionality is enabled by the pre-computation of window information for each nested query. LARC takes each concordance line that features *contributory* as a collocate of *negligence* and stores the individual windows between node and collocate in a list. It then encodes a CQL query in the returned XML data for *negligence* which provides for the largest gap that has been found between the two words.

One feature of an iterative workflow is that the user may encounter dead ends. This can happen where they have drilled down to a certain level of collocations and found that there is no useful information. This gives rise to a need to be able to navigate backwards and forwards through successive queries so that the user is never in a position where they cannot retrace their steps to arrive back in a situation where information is relevant. Many search systems would require a fresh query at this point. However, a design decision was taken in LARC to implement an iterative query history. Every time the user makes a search request during a particular authentication session with the system, the content and parameters of the query are saved to a history menu in the language search interface. Entering the list and selecting any of the listed queries returns the user to the collocate

screen for that particular request.

The design here also seeks to address an issue that was highlighted in the studies from Chapter 4. Respondents found that facilities in existing legal information systems to keep track of sources and information that they had viewed and wanted to return to were lacking. Users often used bookmarking and history tools in web browsers which were said to be sub-optimal, or alternatively to open multiple tabs for individual cases and legislative instruments that could be all too easily lost by accidentally closing their web browser. The language query history in LARC sits alongside the shopping basket functionality that was described in Chapter 5. These way-marking options seek to make it transparent and simple to maintain a record of important information and to ensure that the user is not left in situations where they have to manually correct information deficiencies by starting searches from the beginning. The interrelationship between the shopping basket and the language query history will be discussed further in Section 6.8. The full program code for extracting collocates in response to user queries is provided in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

6.5 From collocations to concordances

Once a user has drilled down to a collocation of interest, the next step is to retrieve a concordance of results. This concordance can either be for the overall query that they have provided or it can be the full set of results for a particular collocation of interest. A toggle at the top of the collocations screen, which is shown in Figure 6.2, takes the user to the concordance. Figure 6.4 shows the concordance results for a simple query on *negligence*. It is also possible to ask the system to return a concordance for a specific collocate which has been ranked as significant in the previous step. Here, the user clicks on the drop-down menu to the left of the collocate that they are interested in and selects the option to *View Concordance*. Using the concordance toggle shows all results for *negligence* whereas the contextual menu option shows all results in the database for the specific collocation, like *contributory negligence*, that has been selected. A concordance is essentially a group of long lines with the query node at the centre. It is designed to show contexts from the database where a particular term arises. By default, LARC implements the standard KeyWord In Context (KWIC) presentation paradigm for concordance lines. However, another way of viewing the data is available and this

will be covered in Section 6.6.

As before, the Manatee corpus manager is agnostic about ranking and filtration algorithms for concordances. This means that any algorithms which are implemented to assist in the viewing of contexts around search hits are specific to the LARC platform. By default, LARC implements no organisation on concordances for either a high-level query or a collocation of interest. This means that the contexts around search hits are presented in the order that they are identified in the corpus. The system employs a paginated layout for the display of results which means that a user can navigate through a concordance to find interesting text. However, an unfiltered and unranked concordance can be difficult to work with because there is no idea of the significance or the frequency of different lines. It also becomes apparent that many results are repeated or duplicated. This tends to occur because much of language is essentially boilerplate. Thus it is necessary to implement some algorithms and facilities in the system to identify important concordance lines and to filter results so that duplication is minimised.

The primary mechanism for organising concordance results in LARC is an algorithm called *Typical*. It scores lines according to the saliency of all the words around a search hit. It also attempts to group lines which feature similar linguistic constructions together. This is intended to provide the user with a concordance view that is easier to read and to navigate. The groups of similar hits should be readily apparent and paging through the results will move between the groupings that have been found for different types of language.

The algorithm works by iterating over each token in a concordance line. First, it extracts the corpus frequency for the word and the frequency of that word in the current concordance. Two factors are then created from both figures, which are the concordance frequency divided by the line size and the corpus frequency divided by the corpus size. The ratio of these factors is then calculated. This ratio is stored to a list of values for every word in the current line. Once all the words in the line have been processed in this way, a mean ranking factor for the line as a whole is calculated by averaging the word factors. Each word value is then squared and subtracted from the mean significance figure that was produced in the previous step. The mean value of this difference figure for each word is produced and stored along with the standard deviation of the difference figures. Next, the z-score for each token in the current line is calculated, which is the raw salience figure minus the mean salience divided by the standard deviation for the

line. Each z-score for an individual word in the line is added together to give a normalised salience measure for the line as a whole. The lines of text for all hits in the concordance are then sorted in descending order of these salience figures for the individual lines. The full Typical algorithm can be seen in Figure 6.12.

$$\begin{aligned}
 p_s &= \frac{f_{conc}}{len_{line}} \\
 p_c &= \frac{f_{corp}}{size_{corp}} \\
 x_i &= \frac{p_s}{p_c} \\
 m_{xi} &= \frac{sum(xi_{line})}{len(xi_{line})} \\
 sq_{x_i} &= (xi_{line})^2 - m_{xi} \\
 m_{diff} &= \frac{sum(sq_{x_i})}{len(sq_{x_i})} \\
 \sigma &= \sqrt{(m_{diff})} \\
 z_{token} &= \frac{(x_i - m_{xi})}{\sigma} \\
 s_l &= sum(z_{token})
 \end{aligned}$$

Figure 6.12: The Typical concordance line significance algorithm in LARC.

Although the grouping of similar lines in a processed concordance works reasonably well under the Typical algorithm, this still means that the user has to

navigate through a large amount of data in order to see examples of different language which feature their query node. In order to ameliorate this problems, LARC provides an option in language search to filter concordances that have been ranked with Typical. This attempts to remove lines of language that are similar to entries already seen in the result set. The filtration is achieved by extending the algorithm shown in Figure 6.12.

Because one step in Typical involves calculating z-scores for each individual word in a concordance line, it is possible to identify the highest z-score for each hit environment. This high score is indicative of the most informative word which has been found in the line. The filtered typical algorithm works by keeping a record of the highest z-score for each line in the concordance. Where a high z-score has been seen before, lines are removed from the concordance until a new z-score (either a higher or lower score than the previous line) is encountered. The intention is to use this metric to remove essentially similar lines of text from the typical groupings so that the user is left with one hit for each different type of language use. A duplicate high z-score is considered to represent the same dominant information word in subsequent concordance lines. The full program code for extracting concordances from the corpus is given in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

6.6 Looking at variation outside the node

The Typical algorithm and its filtered variant which are implemented in LARC help to present key information by first grouping and then attempting to eliminate language information that is essentially duplicated within a concordance. However, it can still be challenging for the user to appreciate how a query term has been treated across the large range of different cases and items of legislation in the database. It is suggested that an appreciation of linguistic variation helps the legal researcher to understand how terms have developed, how they have been constrained and how their definitions have been set in context in different scenarios. There is a fundamental problem with the KeyWord In Context (KWIC) visualisation which is the default mechanism for displaying results in the LARC platform. The listing of results is not very useful for elucidating linguistic variation outside of that node.

As discussed in Chapter 3, a number of different visualisations have been proposed

which seek to augment or replace KWIC. The justification for these developments is sometimes tenuous. However, it is suggested that a class of visualisation which presents concordances as tree diagrams are of utility. Their usefulness stems from the fact that they highlight how a group of search results vary at different token positions outside the query node. If the idiom principle as proposed by Sinclair is a valid theory for how language is built and used to convey meaning, then it follows that some type of hierarchical visualisation becomes appropriate for highlighting language choice.

LARC implements a variation of the *DoubleTree* visualisation for highlighting linguistic variation and idiomatic word choice in any concordance for a given query. Previous iterations of this paradigm have been limited because they present the tree diagram as the ultimate step in a search and retrieval exercise. The biggest problem with existing tree implementations is that they do not preserve a concrete idea of the sentence, or concordance line, as a valid beginning for subsequent focus and search requests. Thus the idea has been expanded in LARC so that, once a user has viewed a concordance in tree mode and has made some linguistic choices at different token positions, they can right click on a node and transfer the partially-built utterance into a new concordance query.

The tree visualisation is implemented in software by loading a standard Manatee concordance - which may either be ranked by Typical or unranked - into an in-memory SQLite database table. The software breaks the concordance lines down into a series of nodes by running `SELECT DISTINCT` queries at each token position in all concordance lines in sequence. This means that, for each token position to the right and left of the query node, a list of the choices of words which are available is returned. Each individual choice is then presented as a node on the tree. Subsequent nodes are computed by moving out to the right and left of the previous token position. It is therefore possible to generate a dependency tree once any number of initial language choices have been made. The resulting tree is interactive and the user can click on any token position in a concordance line to retrieve choices for words in either the left or right co-texts. The visualisation can be seen in Figure 6.13 and the program code for this element is located in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

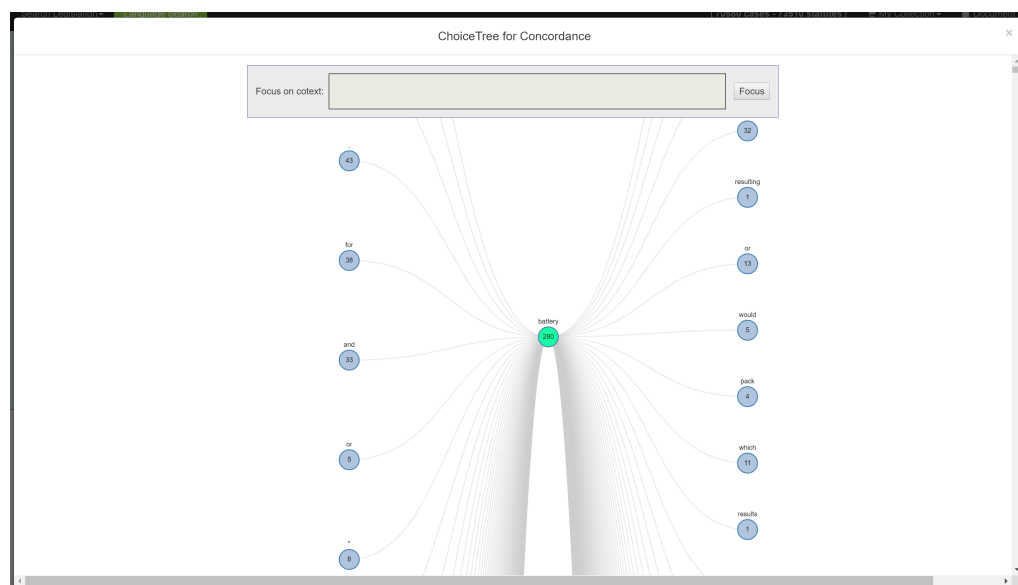


Figure 6.13: The tree concordance for a query on *battery*.

Aside from the tree visualisation for concordances, LARC provides a couple of facilities which make it easier to understand variation in search results directly from the KeyWord In Context display. Firstly, once a query has been run and results returned, it is possible to sort the concordance on any token position in either ascending or descending alphabetical order. This serves to group the same tokens in a particular word position together. The user can now navigate the result set with a focus on the different words which appear alongside other tokens at a particular position in the line. Another facility which has been implemented is called *ad-hoc classes*. The rationale here is that, whilst the software can do a certain amount to isolate important language for the user, it is the lawyer who is the ultimate arbiter of relevance. Thus the ad-hoc classes dialogue allows a concordance to be filtered progressively on word position. The user may be interested in all instances from the corpus where *contributory* appears one position to the left of the *negligence* node. They can specify this word in an ad-hoc class request which can be passed a parameter either to keep only qualifying co-texts or to discard them from the concordance. Ad-hoc classes can be implemented in a progression to filter the concordance according to information need as the language around a query becomes more apparent to the user.

One limitation of both the traditional KWIC layout and the tree-based diagram for visualising concordance content is that the length of individual result lines, and therefore the amount of information about a search hit which is communicated to

the user, is limited by the horizontal size of the result and, ultimately, the display. The XML data which is generated from query requests for concordances includes information about the position of each search hit within the corpus as a whole. This positional information allows for a *full co-text* display which presents much more content from the environment of particular search hits. In order to retrieve a full co-text for a given concordance line, the user can click on the contextual menu next to the line and select an option to *View full co-text*. This choice opens a drop-down panel at the bottom of the screen. The panel contains a broad window of text around the search hit with the query node highlighted. A full co-text view for a particular search hit on *negligence* can be seen in Figure 6.14.

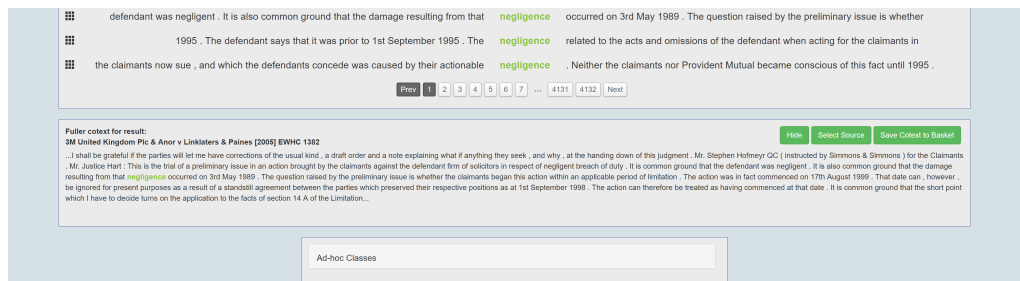


Figure 6.14: The fuller co-text for a search hit on *negligence* from the returned concordance.

6.7 An infrastructure for caching results

Most corpus interrogation systems implement protections against long-running queries. This class of information request features words which are very common in the corpus as a whole. The primary mechanism for ensuring that a system is responsive and delivers results promptly is the *stop list*. A *stop list* is a collection of tokens or collocations which are seen very frequently and which the user is prevented from searching for. As an example, a query for the word *a* on the case law corpus in LARC, which consists of some 400 million words, takes just over an hour to return results. Clearly such a timeframe is problematic for the user experience. Very frequent queries also tend to be computationally intensive. However, the use of a *stop list* is not ideal. By preventing certain queries from being run, the system designer is saying that technical overheads dictate platform operation and prevent what may be legitimate interrogation requests on the database.

A design decision was taken to remove the need for *stop lists* in LARC. Implement-

ing this feature relied upon the pre-computation of collocate and concordance result sets for frequent queries. The underlying information response to a request in language search is an XML document which is then parsed by the view layer of the system. Thus it was possible to generate XML documents for frequently-occurring tokens in advance. Both the collocate and concordance routines take user input as a search request and then find the numerical identifiers of the constituent words in the corpus. Each unique token in the corpus has an identifier which is an integer value. The system then looks on disk storage for a folder which is named with this identifier. If a folder is found, the software searches inside it for a file that features the unique identifier of a particular collocation. A live query for that word is therefore avoided and query times are virtually instant.

In order to facilitate the pre-computation of common queries, a parallel federation of five dedicated servers was created. These resources were installed with RabbitMQ, a distributed messaging service. Message queues were created in the RabbitMQ system for collocation and concordance queries. These queues were first loaded with a randomised list of the most common words from both the case law and legislation corpora in LARC. Randomising the queue contents allowed each server to receive a computation request, to generate the XML required and then to move on. One or two of the servers may have been unavailable for a long period whilst working on a very frequent word but there were others available which were not blocked.

The processing of individual queries was done by a pool of worker scripts on each of the available servers. A worker waits for a message to be available on a given queue, retrieves it and then hands processing off to a background script. The worker is blocked for incoming messages until the current computation has been completed. Once the XML file is generated, the worker acknowledges the message - which removes it from the queue - and waits to receive a new request. It took a total of four months to process all five hundred thousand of the most common words in case law and legislation corpora for LARC. The RabbitMQ environment is shown in Figure 6.15. A sample worker script for pre-computing collocations can be seen in the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

6.8. Tying it all together - the complete interface

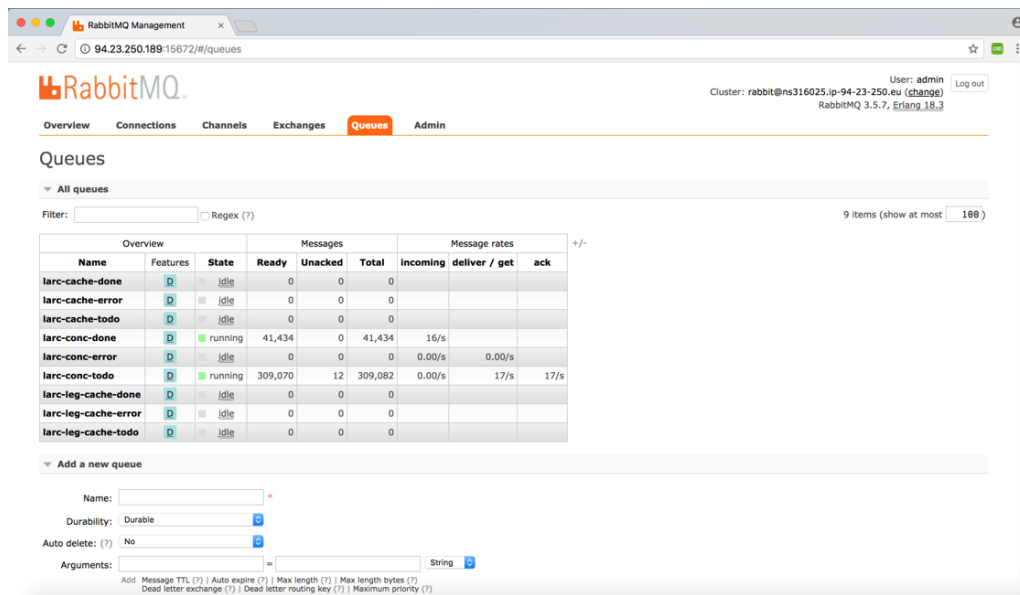


Figure 6.15: The RabbitMQ federation manager used to precompute frequent queries in LARC.

It was initially intended that all single-word queries against the corpora should be pre-computed. This would then be augmented with all first-level collocates for those tokens. This meant that all results for *negligence* would be available instantly as well as all results for *contributory negligence* and, indeed, for *negligence* immediately proceeded or followed by any other token. However, experiments showed that this resulted in a message queue of some twelve million queries to precompute on each corpus. The available time and processing resources dictated that only single root queries were ultimately cached.

6.8 Tying it all together - the complete interface

Language search is presented as an additional facet in the information-seeking screens of LARC. A user launches interrogation requests against the case law and legislation corpora simply by selecting *Language Search* from the main program menu. The search screen features a query box into which the user types words (or a phrase) that they are interested in. It is possible to constrain a search simply to items of case law or legislation which are stored in the current shopping basket (see Chapter 5), or to run a query against the whole corpus. The case law and legislation corpora are segregated such that the user must choose which collection they wish to search against before submitting a query. This was done partly for

practical reasons and partly because the findings of the survey work in Chapter 4 indicated that search results from multiple sources in the same set are confusing and difficult to deal with. Before sending a request, the user can specify whether they want to perform a search for all variants of a root word (a lemma search) and what number of intervening words should be considered between the tokens of a query in the results. They may also choose from any of the collocate ranking algorithms presented here.

As soon as a result set is obtained for a language query, LARC presents a collapsible *Key Sources* panel at the top of the screen. This shows those cases or items of legislation which most commonly feature the given query. The table of sources can be organised by citation index or by result count simply by clicking in the appropriate header row. If the user wishes to research a particular source, clicking on its title in this panel redirects the system to *Research View* and loads the appropriate report or act text. In collocation mode, a contextual menu positioned to the left of each entry in the list of tokens allows the user to see cases or statutes which feature the given collocation. Clicking on any of these instrument titles redirects the system to *Research View*. The facility to drill down into progressively more detailed collocations is also accessible from this menu.

In concordance mode, a contextual menu to the left of each concordance line allows the user to retrieve a full co-text for any individual result. They may also select the source of a particular line, which redirects the system to *Research view* and loads the appropriate case or act text. Facilities for sorting the concordance by token position, for extracting a tree-based visualisation of the concordance and for exporting the concordance to either Microsoft Word or PDF format, are also available at the top of the screen. The ad-hoc classes functionality occupies a collapsible panel at the bottom of the screen, where filtration requests on the concordance as a whole can be implemented. Both collocation and concordance views integrate with the session and shopping basket facilities which are described in Chapter 5.

6.9 Algorithmic evaluation

In order to evaluate how well the different algorithms that have been proposed in this chapter for language saliency calculation and ranking work, test runs of each different approach are now presented. The top twenty collocates for a query of

negligence are shown as returned by each different algorithm.

It should be noted that the score associated with each ranking for an individual collocate is essentially arbitrary. The important thing about the scores is their relative magnitudes which denote how strong the relationship between a collocate and the node word (*negligence*) is for each entry. In general, a tighter scoring with less significant differences between positions in the list is preferable because it allows for an easier presentation of the importance of individual rankings. The LARC platform exposes the scores for individual collocates when the user hovers over the contextual menu to the left of an individual word in collocation view.

For concordances, a result set for the word *hot*, which has many distinct meanings associated with it, is presented under the *Typical* and *Typical-with-filter* approaches. See Figure 6.16 and Figure 6.17 respectively here.

6.9.1 Collocates of *negligence*

The results of the observed and expected algorithm for collocate ranking on a query of *negligence* can be seen in the column for Algorithm 1 in Table 6.2. Although the ranking and content of the process is reasonably informative, there is a tendency for unexplained elevation of non-frequent words like *Copleys* and *Knightly*.

The results of the z-score algorithm for collocate ranking on a query of *negligence* can be seen in the column for Algorithm 2 in Table 6.2. This approach works well. Each of the ranked collocates is informative and the order of the list is plausible.

The results of the t-score algorithm for collocate ranking on a query of *negligence* can be seen in the column for Algorithm 3 in Table 6.2. The results do not feature any words which have been elevated in significance for no apparent reason. However, frequent closed-class words with little informational value are ranked consistently highly. These results are not as useful as those returned previously for *z-score*.

The results of the *Mutual Information* algorithm for collocate ranking on a query of *negligence* can be seen in the column for Algorithm 4 in Table 6.2. It is immediately apparent that these rankings are equivalent to the observed and expected approach. The only difference is that the scores associated with each collocate are of a different order of magnitude. This is because the logarithm in the equation is fixed and so the algorithm is essentially identical to the default approach.

1. Observed/expected		2. Z-score		3. T-score		4. Mutual Information		5. Log likelihood	
Collocate	Score	Collocate	Score	Collocate	Score	Collocate	Score	Collocate	Score
contributory	1326.94828974	contributory	56.398246957	in	146.663251643	contributory	10.3738964356	in	192473.633965
Contributory	774.708246158	negligence	18.4596739952	a	123.382591622	Contributory	9.59750928588	a	129463.432421
discounenance	742.556322899	gross	13.9408031357	is	104.779763051	discounenance	9.53635664826	is	89735.2492695
Deficient	703.244517569	caused	10.6870676023	was	97.0337462118	Deficient	9.45788259024	was	76913.3878078
Honesty	668.300690609	liability	10.531375263	or	93.8061718858	Honesty	9.38435355482	contributory	70544.9784336
Deliberate	465.785329818	Contributory	9.97679634616	not	89.249452731	Deliberate	8.86352139151	or	70124.142683
indolence	465.785329818	damages	9.40245086781	by	87.6920744099	indolence	8.86352139151	not	63445.1378212
misconducts	412.246786161	professional	9.11078444779	on	82.4616501293	misconducts	8.68736443636	by	60855.7397785
untechnical	376.738134412	duty	9.05539759403	as	77.3502210048	untechnical	8.55741826379	on	53941.7864227
vituperative	310.523553212	breach	8.9993759161	be	75.0147996	vituperative	8.27855889079	as	47494.6673012
crass	295.594536231	in	8.96678447952	claim	73.0491481425	crass	8.20747579273	claim	46720.3558388
inadvertence	268.418664641	claim	8.42078556415	it	69.6645751883	inadvertence	8.0683411834	be	45553.9543991
rashness	260.523998034	damage	8.38436755816	have	64.4587382396	rashness	8.02527246151	negligence	44839.8344555
chamfer	254.36503827	nuisance	8.09183881077	contributory	63.9829878622	chamfer	7.99075657981	it	39038.155695
Negligence	245.663982451	liable	7.53986105049	case	62.29143804	Negligence	7.94054254504	breach	34633.8186703
Copleys	243.98279181	tort	7.50808935623	negligence	61.0555225675	Copleys	7.93063558737	liability	34509.7623444
Instances	239.103135973	a	7.15565314821	breach	60.4775291932	Instances	7.90148924171	duty	34416.1196404
Knightly	238.308773395	negligent	7.12705084918	duty	60.1758801631	Knightly	7.89668825545	have	32462.8355064
Fault	229.416654985	action	7.03204234527	liability	58.4492917234	Fault	7.84182632041	damages	31250.1954823
vicariously	221.802538009	clinical	6.99477037404	damages	56.3805327973	vicariously	7.79313206362	caused	30850.207716

Table 6.2: The top twenty collocates for *negligence* under the **observed/expected (1)**, **Z-score (2)**, **T-score (3)**, **Mutual Information (4)** and **Log Likelihood (5)** algorithms.

The results of the *log likelihood* algorithm for collocate ranking on a query of *negligence* can be seen in the column for Algorithm 5 in Table 6.2. There are no falsely-elevated words in the result set. However, common tokens with relatively little information value are again prevalent towards the top of the rankings. These results tend to show that *z-score* is the most robust ranking algorithm for collocates whilst the decision to split open class and closed class words in the LARC interface is justified.

6.9.2 Typical concordance for *hot*

Figure 6.16 shows an extract of a *Typical* concordance for the query *hot*. The result shows that groupings of similar language are extracted quite effectively. The groupings do not tend to be completely coercive or strong, however. The high *z-score* of a line can dictate, for example, that new words like *briquette* can form a new grouping when logically the line should belong to the same group as *rolled steel*. The filtered *Typical* concordance for the same query of *hot* is shown in Figure 6.17.

6. INTEGRATING LANGUAGE SEARCH

Left context	Node	Right context	Score
H5a H5bk) . The following was said by F : H5g line 6 the	hot	tap H5g line 11 in the bathroom and it was right on the top H5x	95.081950405
the top H5x line 22 he was just turning the tap on when it was	hot	H5y line 29 it was on purpose H5sz line 19 24 DC Taylor you told	62.127294462
, Pilatre de Rozier , a scientist , made the first recorded launch of a	hot	air balloon called " Aerostat Revolution " . It contained a sheep , a duck	54.000565319
heading " Bathroom Hot Water System " that the complaint is the inability to have a	hot	shower . The words so as to make them hot- needed to be added to	49.6396833781
like I was washing my hands and he just comed sic in and turned the	hot	water H5z line 27 Okay what happened to your feet ? H5z line 28 29	45.337258713
a stage coach the steersman and the rowers of a boat the workman who draws the red	hot	iron from the forge , and those who afterwards hammer it into shape the engineman and	42.75452937
defects in the donkey engine pipeage , could not be from fresh water or the	hot	well , the learned Lord Ordinary thus treats that subject it is admitted that when the	41.80517512
meatball marinara was transferred to another container , (a Bain-marie) , in the	hot	well of the sandwich counter unit . The temperature of the meatball marinara was allowed	38.952718956
of the premises for the preparation of food on the premises both for re-heating and	hot	from the oven . Such food included not only lasagne , pastas and risottos etc	38.3838309781
vessel having started with her fore feed-pump defective , and the donkey-engine pipeage to the	hot	well gone , the pumping power of the donkey-engine had to be requisitioned somehow to	38.029290937
correct operation between air temperatures of +5 (C and +40 (C . For very	hot	environments (e.g. hot climates , steel mills , paper mills) extra requirements may	37.896710664
temperatures of +5 (C and +40 (C . For very hot environments (e.g.	hot	climates , steel mills , paper mills) extra requirements may be necessary " .	37.52300675
based dishes , toasted sandwiches , tea-cakes , baguettes and croissants . The preparation of	hot	food shall include the reheating of cold food by microwave oven . The baking of	37.387606743
because : (1) Mrs Mulligan deliberately kept the meatball marinara hot in a	hot	well at temperatures of 63 to 68 degrees centigrade after the marination process had been	33.5811695891
completion of the cooking process . (10) Mrs Mulligan sold the meatball marinara	hot	. (11) Mrs Mulligan could not sell the meatball marinara in a hot	32.7748213408
ambient air temperature because : (1) Mrs Mulligan deliberately kept the meatball marinara	hot	state . (9) Mrs Mulligan deliberately kept the meatball marinara hot after completion	34.8942153117
freshly prepared products which could only be achieved if the meatball marinara was in a	hot	condensers and stoves- 19 is not of itself physical or material damage but is consequential on	31.9970791187
possible . That notwithstanding , Mr. Field persisted in submitting that the cost of keeping	hot	tear cracking during solidification of the welds and it may well migrate to grain boundaries	31.9856499056
work-piece and de-stabilises the arc . Any significant lead contact will give the probability of	hot	metal in the shovel . Mr Smith started on the driveside side of the primary	31.2758397222
an oxygen-propane cutting flame to remove the external portion of the bolts , catching the	hot	surfaces inside cars ; and cooling enclosed spaces . At page 2 lines 45-47 the	30.9419836663
sunstroke , heatstroke , burns and sunburn ; treating fevers and hot flushes ; cooling	hot	flushes ; cooling hot surfaces inside cars ; and cooling enclosed spaces . At page	30.690745904
after exercise ; treating sunstroke , heatstroke , burns and sunburn ; treating fevers and	hot	works . All cargo tanks and slop tanks must be cleaned , sludges removed and tanks	30.4633321768
and accommodation spaces must be cleaned and gas freed so that they become safe for	hot	works . It has not been possible to ascertain the exact amount of sludges that	30.294296004
cargo tanks and slop tanks must be cleaned , sludges removed and tanks be prepared for	hot	side bypass system in order to mitigate the problem of VOC-laden air overheating . However	30.1687290756
quality making steel products by way of electric furnaces , especially fat products such as	hot	rolled coils . The direct reduced iron is mainly used by Megateel , an associated	29.8133408303
a condition known as proctagita fulgax which has been likened to have having a red	hot	poker thrust up the anal orifice . For as long as it lasts she is	29.499113574
what BRIAN RAWLING PRODUCTIONS is . It means clarity - RIVE DOITE MUSIC is a	hot	publishing company and BRIAN RAWLING INC Productions is what it says it is . I am	29.0597476017
box at the bottom it is stated that : " This contract is valid for	hot	of as and/or brochure publication alike . " Then , under the heading " Comments/Additional Clauses/Special	28.9880068
, a well known and very reputable local firm . ENTRANCE HALL ; cupboard with	hot	water tank . BATHROOM ; with shaped corner bath , jacuzzi and mixer tap with	28.699961528
(Day 2/151-2) : " 10) new grain etc - 2x heaters replaced	hot	water tank as old were down to earth and taking out rcd : completed .	23.020019762
Gujarat and listed on the National Stock Exchange . Its business is the manufacture of	hot	bricquetted iron , hot rolled steel coils and other similar products . On 22nd March	22.8911229305
the National Stock Exchange . Its business is the manufacture of hot bricquetted iron ,	hot	rolled steel coils and other similar products . On 22nd March 1999 it was due	22.564307808

Figure 6.16: A sample Typical concordance for the query *hot*

Left context	Node	Right context	Score
H5a H5bk . The following was said by F : H5g line 6 the	hot	tap H5g line 11 in the bathroom and it was right on the top H5x	95.081590405
, Pilatre de Rozier , a scientist , made the first recorded launch of a	hot	air balloon called " Aerostat Reveillon " . It contained a sheep , a duck	54.6005663919
meathall marinara was transferred to another container , (a bain-marie) , in the	hot	well of the sandwich counter unit . The temperature of the meathall marinara was allowed	38.9532718956
sunstroke , heatstroke , burns and sunburn ; treating fevers and hot flushes ; cooling	hot	surfaces inside cars ; and cooling enclosed spaces . At page 2 lines 45-47 the	30.6900745904
replacing the old polymer dryer with a new dehumidifying dryer . He explained , "	hot	air dryers will not dry polycarbonate efficiently . Whilst a dehumidifying drier will remove moisture	24.0325601328
Gujarat and listed on the National Stock Exchange . Its business is the manufacture of	hot	briquetted iron , hot rolled steel coils and other similar products . On 22nd March	22.8788883415
repair yard or scrap yard . " In the DNV guidelines on lay-up , "	hot	lay-up " is defined : " In this lay-up condition , the machinery is kept	21.4261474966
: 2.7 g (8.2 mmol) of paroxetine was dissolved in 15 ml of	hot	ethanol : 1.0 g (10.4 mmol) of methanesulfonic acid in 15 ml of	20.1702947377
. (11) Although the adverts for toasted Subs did not use the word	hot	, the Tribunal found that the strap-line of fresh toasted and the images of browned	16.2412301329
discuss a variety of uses for a water-based cool mixture : cooling the body during	hot	weather or after exercise : treating sunstroke , heatstroke , burns and sunburn ; treating	15.2090817819
The condition does not define the term " hot snacks " . By definition a	hot	snack could well be defined as a hamburger and chips or a hamburger . A	13.9494781891
Ivess presence . (6) Mr. Ivess opinion as to the temperature at which	hot	drinks ought to be served in McDonalds restaurants . (8) Comments on the	12.8407377766
is very similar to the events of 12 February 2006 , when Mr.Lemon drained the	hot	water tank without switching off the heater . Mr.Lemon did not give evidence but ,	12.4051000974
Wilde had said when cross-examined about the implications of not keeping the condensers and stoves	hot	was that the damage to the condensers and stoves , had that not been done	11.9381741992
the molten plastic in the sprue on the other , and the distinction between a	hot	sprue and a cold sprue lies in the location of that region of transition .	9.87968544647
as part of an agreement that Paramount would be a distributor for Hotspring spas and	hot	tubs . He said that Mr.Biggs of Watkins had emailed the Hotspring UK database	9.12090101212
119 , as follows : The phrases " approbating and reprobatng " or " blowing	hot	and blowing cold " are expressive and useful , but if they are used to	8.73324345078
, forced exertion such as forced running , temperature manipulation such as detention in unbearably	hot	locations or dousing with cold water and sensory bombardment or use of noise) ;	7.93586512048
the plumbing which connects the hot and cold water taps on a bath to the	hot	and cold water tanks in a house . The connections are entirely common-or-garden but the	7.91129758844
Mr.Palfrey is recorded as having told Mr.Boyle this : " 6 pm accidentally switched off	hot	feric heater . Washing castings near the anodic panel . Think Andy might have reported	7.34348761635
Mrs Conradi 's evidence that Mr. Noble would find that a Tempur mattress makes him	hot	and sweaty Tempur overlay Mrs Ho was of the view that this item would assist	6.87193456044
body 2 being removed for heating and then propelled onto the food . Said propelled	hot	air cools in contact with the food and is removed again for reheating , and	5.44881921249
a highly combustible rubber material that was also red . He said that carrying out	hot	work in the immediate vicinity of the Linotex was a " no-no " . The	5.42680808004
red , hot and swollen , also exquisitely painful . Left lower limb swollen ,	hot	and painful Left leg still feels numb over thigh Could not feel dorsalis pedis The	4.83592531571
fire . (4) Neither the water being sprayed into the underpan whilst the	hot	cutting work was taking place nor the water used when the underpan was being hoised	4.14173123541
the wire in the same way as wires were operated in hot biopsy forceps and	hot	snares : Dr Williams (Olympus 's expert) recognised the handle shown and described	3.57420793282
and high shrinkability soils forming the Brickearth and Reading Beds geological strata . During the	hot	, dry summers that occur in southern England , on average every five years ,	3.29569481482
pressed hot water passes over the grounds and into a jug which rests on a	hot	plate keeping the brewed coffee at the desired temperature . Since 1993 , the coffee	2.87239602066
the Company Alplasteed uses land which is owned by Technoplan Anstalt , Liedtstein . A	hot	strip mill used by the Company is owned by Lictor Anstalt , Liedtstein . There	2.75834738563

Figure 6.17: A sample filtered Typical concordance for the query *hot*

6.10 Conclusion

This chapter has introduced the linguistic search facilities offered by the LARC platform. This work forms **Contributions C2.1 and C2.2** of the thesis, as enumerated in Chapter 1. The approach taken here is based upon algorithms and interface design norms from the domain of corpus linguistics. The methodologies have been used to create a guided search mechanism which is based upon presenting collocations as a fundamental unit of meaning. The idiom principle, as proposed by John Sinclair, has been used to justify this method of interrogating language use.

The resulting databases are designed to be living resources. This means that they can be updated over time as new legal information is published without significant technical or practical overheads. The design of the corpora is based on a decision to use plain text in their construction as far as possible. This removes problems which can arise when using sources that have been heavily augmented with derived streams of information.

There is a certain tension between this desire for plain text and the need in LARC to tie language back to sources which are organised and faceted. The underlying query language used by the Manatee system - Corpus Query Language or CQL - also imposes some structure on the sources used, as well as introducing complexity and limitations on the prospective users of a system. LARC takes advantage of a largely plain text data source in order to hide query complexity from the user in a simple interactive interface.

LARC seeks to use the corpora to build an iterative search experience which concentrates initially on refining information need based upon drilling down into interesting collocations that are presented for a root query. Different algorithms for extracting these collocations from a corpus on the basis of their saliency in relation to a given node word have been presented. A design decision has been taken here to treat the available corpora as large enough and broad enough to be representative of language in the legal process. Thus metrics used in some of the algorithms (particularly Mutual Information) that are traditionally built on probability calculations have been replaced with alternatives based on observed frequency measures.

LARC implements a segregation in the presentation of collocation search results between open and closed class words. This has been done to prevent search results

being polluted by highly-frequent words which do not have large information value. The evaluation results from the different collocation ranking algorithms indicate that most approaches suffer from elevated rankings for these this type of word in the absence of a separation strategy. The search interface also places important collocates to the right or left of the node word, as appropriate, which is a layout that most other corpus systems do not use.

The workflow presented in this chapter envisages that a user will drill down to an interesting collocation and then switch to the concordance view for the given result set. A concordance uncovers more information about the dominant environments of a query. There is a query history implementation in the software which should mean that users never encounter dead ends in their search requests. A concordance on the system can also be returned either for a root query or for a particular collocation.

Concordances are presented by default in the traditional KeyWord In Context (KWIC) display format. Concordances in this view can be sorted alphabetically by token position and they can also be filtered through the adoption of ad-hoc classes. This KWIC approach is not ideal, however, for looking at linguistic variation outside the node. The idiom principle suggests that a hierarchical visualisation can be used to better effect for this purpose. LARC therefore implements a variation on the *DoubleTree* layout which differs from the canonical version by preserving an idea of the sentence.

Several test runs of the collocate and concordance ranking algorithms that have been described are provided. All the approaches deliver data which is relatively interesting and valid. However, some of the algorithms have a tendency to elevate infrequent words to the top of a result set. Others feature closed class words with little information value heavily. The best starting point for ranking collocations is found to be the z-score metric. The Typical algorithm is introduced for ranking concordances. This attempts to group result lines according to common or essentially duplicate meaning. The results are reasonably effective and the addition of a simple high *z-score* filter allows duplication to be removed well.

IV

PART IV

EVALUATION AND DISCUSSION OF RESULTS

SYSTEM EVALUATION

7.1 Thesis process

This chapter contributes to answering the main research question in this thesis by addressing the following steps of the process outlined in Table 1.2:

- **P6 - Consider how usable the LARC software prototype is according to standardised user experience metrics.**
- **P7 - Consider the feedback about LARC from lawyers and law students which should form the basis of future work on the platform.**

These steps of the process will be addressed by applying statistical measures of system usability which are provided by standard user experience scales and metrics. Free-form reaction data about the user experience is also reported and analysed. A particular focus on the language search functionality in LARC will seek to reveal the strengths and weaknesses of the information retrieval approach that has been taken, which is directly relevant to the main research question. In Chapter 8, the results from this evaluation will be considered in relation to the original productivity barriers which were identified through a contextual inquiry, interviews and survey work with lawyers and law students in Chapter 4.

7.2 Introduction

This thesis is based upon an initial exposition of user experience and productivity barriers with current legal information systems which may lead to research skills

deficits in newly-qualified and early-stage professional lawyers and law students. It has been found that existing platforms for legal information retrieval and a jigsaw of systems which facilitate legal research at the moment either present usability barriers themselves or fail to answer problems which can impact upon the quality of legal research. LARC is specifically intended to answer the major usability barriers that were found in the contextual inquiry, interview and survey work, based on a consolidated list of observed problems which is presented in Appendix A, Section A.8.

The key focus of this prototyping and development work has been on producing an integrated legal research platform which provides for information discovery, analysis and knowledge synthesis in a single coherent application. The LARC system has been prototyped in paper and in software on the basis of an iterative development effort which uses open source software components in a tightly-integrated framework. The approach here is directed by the technical success of the PhraseBox project [77] (where the corpus manager was released as open source) and other projects described in Section 1.5.1. The resulting software will now be evaluated by a group of law students and lawyers.

The goals of this evaluation are: to find out how usable the LARC system is; to identify shortcomings which exist in the software and which create critical incidents for users so that they cannot proceed effectively; to collect general information about the emotional responses of different users to the way that LARC works; and to find out how well the new approach to language discovery is received. Finally, it is desirable to understand whether different users prefer LARC to other existing tools for legal research.

7.3 Identifying the evaluators

In order to identify participants for the evaluation exercise, an initial email was circulated to every law school and university law department in England and Wales, and to those in Scotland which offered either single or joint honours degrees in English law. For each law school, a senior member of staff responsible for mooted activities or learning and teaching was manually identified from the institutional web page. These staff members were the recipients of the first email about the evaluation. This same email was also sent in a modified form to contacts that had been developed during the course of the research work in several law

firms.

Once an agreement in principle to participate in the evaluation exercise had been received from several law schools and law firms through each senior contact, an online form for participant sign-up was created and published. The law schools and the legal firms then sent the link to this form around members of staff, students and lawyers who might be interested in evaluating LARC. From the responses to this form, an initial pool of 18 participants in the evaluation was collected. These included people who were law students, participants in mooting, members of academic staff and professional lawyers. Table 7.1 provides a breakdown of the proportion of responses by role.

Role	Proportion of responses
Undergraduate law student	55%
Legal academic	11%
Solicitor	29%
Other	5%

Table 7.1: Initial breakdown of evaluator interest by job role

A cut-off date for receiving evaluation sign up requests was set. After this time, user accounts for all of the enrolled participants were created on the LARC system. This meant that they could perform advanced tasks such as creating and deleting documents, navigating the session history archives, exporting the contents of documents and chat rooms to Microsoft Word or PDF and so on. On the creation of each new user account, an automated email is sent to the participant with a secure password which grants access to the system in combination with their email address. Of the sixteen individuals who expressed an interest in evaluating LARC, nine actually went on to complete the evaluation process. Table 7.2 breaks down the data on participants who completed the evaluation exercise by job role. This relatively small conversion rate reflected an ongoing challenge to engage people who had expressed an interest in the evaluation sufficiently in order for them to complete the exercise in full. Nevertheless, nine participants meets a basic and long-standing metric for Human Computer Interaction system evaluation efforts. The metric is defined by Nielsen when he suggests that a cohort of between three and five users tends to identify 80% of usability issues with the system being tested [138].

Role	Number of responses
Undergraduate law student	4
Legal academic	1
Solicitor	4
Other	0

Table 7.2: Final breakdown of evaluators who completed the exercise by job role

7.4 Evaluation task

The objective of the evaluation exercise was to allow participants to use the LARC system as freely and naturally as possible whilst collecting detailed quantitative and qualitative feedback about their user experience. The evaluation itself was designed as a rapid empirical study which was conducted remotely, outside a lab-based setting, over the Internet. The advantages of a rigorous, lab-based study in this context were incompatible with the evaluation pool and their level of engagement in the exercise. The process for running the evaluation, collecting data and analysing that data, therefore, essentially follows the rapid, remote approach which is suggested by Hartson and Pyla in [71].

The LARC system is publicly available and it has been specifically designed and tested for use in Google Chrome. The evaluation was designed such that the participants could perform their evaluation at a time which suited them, in one session or across several engagements with the platform, and in an environment where switching between the evaluation and confidential client work did not present privacy and security concerns. It was desirable to record these sessions in video in order to enable the identification of critical incidents with the software, which will be discussed in Section 7.6. However, in-place recording during sessions was not practical with multiple remote respondents and so a server-side approach to capturing video session data was implemented.

It was desirable to allow each participant to use LARC in their own way whilst ensuring that each discrete element of functionality in the system could be tested appropriately. In order to facilitate this, a design decision was taken that a seed question involving various legal research tasks would be provided to each participant in advance. The question would necessitate usage of different elements of the LARC platform in order to produce a structured response document which contained the outputs of the legal research tasks involved. As a central element of

the LARC platform is the document editor, it was possible for each respondent to compose their work products directly in the system and for the research team to see these results for analysis at a later date. The seed question presented a standard legal scenario of the type commonly given as an exam question in an undergraduate law degree. The question was as follows:

Question

One day, while walking home, William trips and falls, damaging his knee. Several days later, while driving to work, he sees Victor crossing the road and brakes to avoid running into him. Unfortunately, due to the pain in his knee, he cannot fully press his brake pedal and, as a result, he runs into Victor. The collision occurs at a fairly slow speed and a normal person would only have suffered bruising as a result, but Victor has brittle bones and suffers two broken legs and a number of broken ribs. He is taken to the local hospital where, due to an administrative mistake, his right arm is amputated.

Advise Victor.

Instructions

- *Log into LARC*
- *Use the username and password for the LARC system that you have received by previous email.*
- *Find an initial seed case on the system and create a new document for your work.*
- *Use LARC freely to research the problem question.*
- *Place relevant case citations, statutes and language into the document that you have created.*
- *Write an outline answer to the question in the LARC document.*
- *Once you have finished answering the question, please complete the following questions in this survey.*

The problem question provides a scenario which involves addressing breach of a duty of care, the rules of causation and the egg-shell skull principle, a break in the causal chain with a possible intervening act when the hospital amputates the arm and *res ipsa loquitur* in relation to the liability of the hospital. The entire question falls within a branch of English law called *tort law*.

7.5 Metrics used

An online questionnaire was published and an anonymous link was distributed to all the participants who had registered an interest in testing LARC. No identifying information was collected in the evaluation survey save for the job role of the individual participant. This information enables statistical analysis of responses according to the sector, position and seniority of the informant. The evaluation questionnaire was split into two parts.

The first part was designed to collect quantitative data about the user experience in LARC. This portion of the survey used ten questions based upon the standard *System Usability Scale* framework [23]. This ranking and scoring system was chosen because it is easy to administer to remote participants. It has been shown to produce reliable usability results on small sample sizes and its validity has been repeatedly tested in order to ensure that it can differentiate between usable and unusable systems. The quantitative element of the questionnaire thus involved questions about the usability of LARC which required a ranking in each case on a five item Likert scale, from *Strongly agree* to *Strongly disagree*.

The rest of the survey involved free text entry responses from each participant. These questions focussed on emotional and overall practical reactions to the system. They required each respondent to detail what they liked most about LARC; what they disliked most about the system; what they would change in the platform to make it more useful in their work; how useful the language search facility was; and what they felt were the best elements of the system for training law students. The full questionnaire that was used in the evaluation can be seen in Appendix C, Section C.1.

7.6 Data collection approach

Aside from the evaluation questionnaire, screen capture videos were recorded from each evaluation session. This was done in order to be able to identify critical usability incidents with the platform which could inform future improvement work. These critical incidents are described and discussed in more detail in Section 7.7.3. A server-based solution called *Hotjar* was identified for recording. This platform operates through Javascript and it captures the entire webpage in which it is instantiated but recordings are limited to that page. Thus it was possible to record evaluation sessions in LARC simply by adding a block of Javascript code

to the shared header of the two main layouts in the application, search facet view and research view.

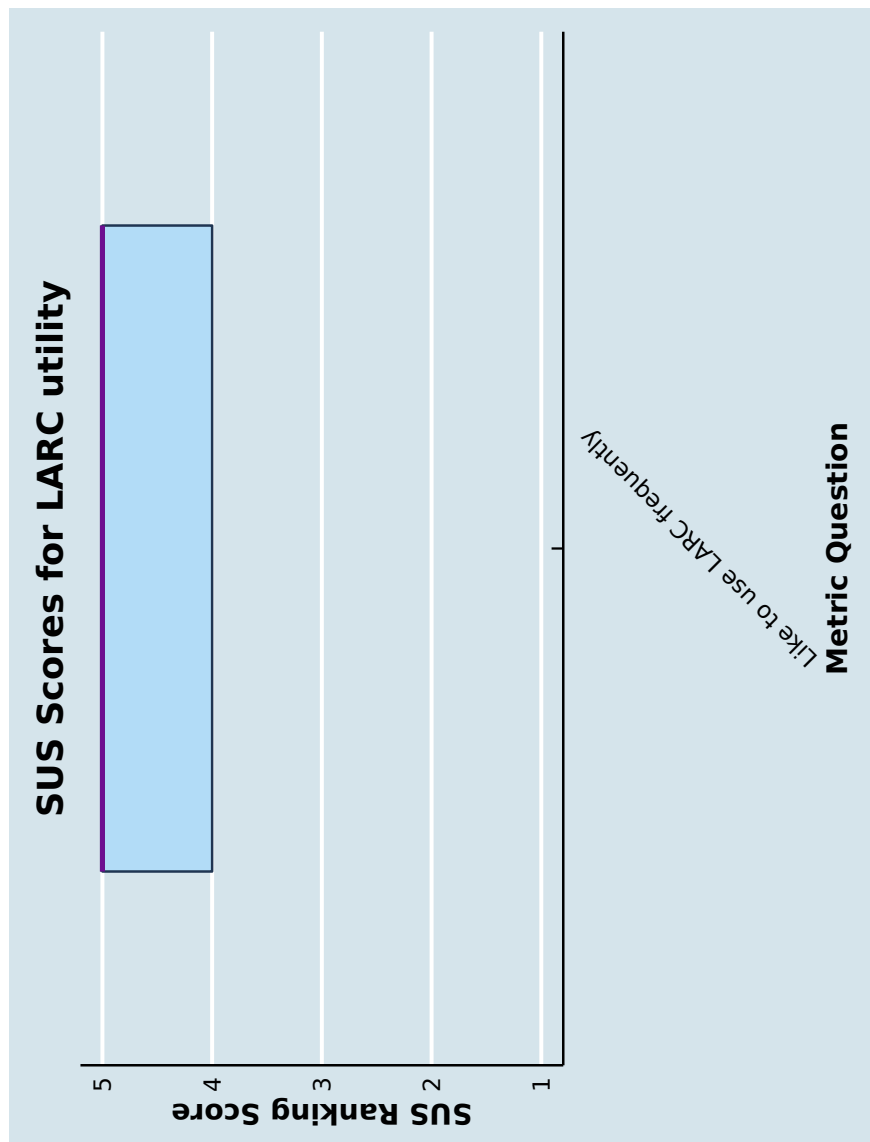


Figure 7.1: Scores from the evaluation participants on SUS questions relating to the utility of the LARC prototype. The scoring scheme is a five point Likert scale from 1 (strongly disagree) to 5 (strongly agree). See **Section 7.7.1** for an explanation of the diagram.

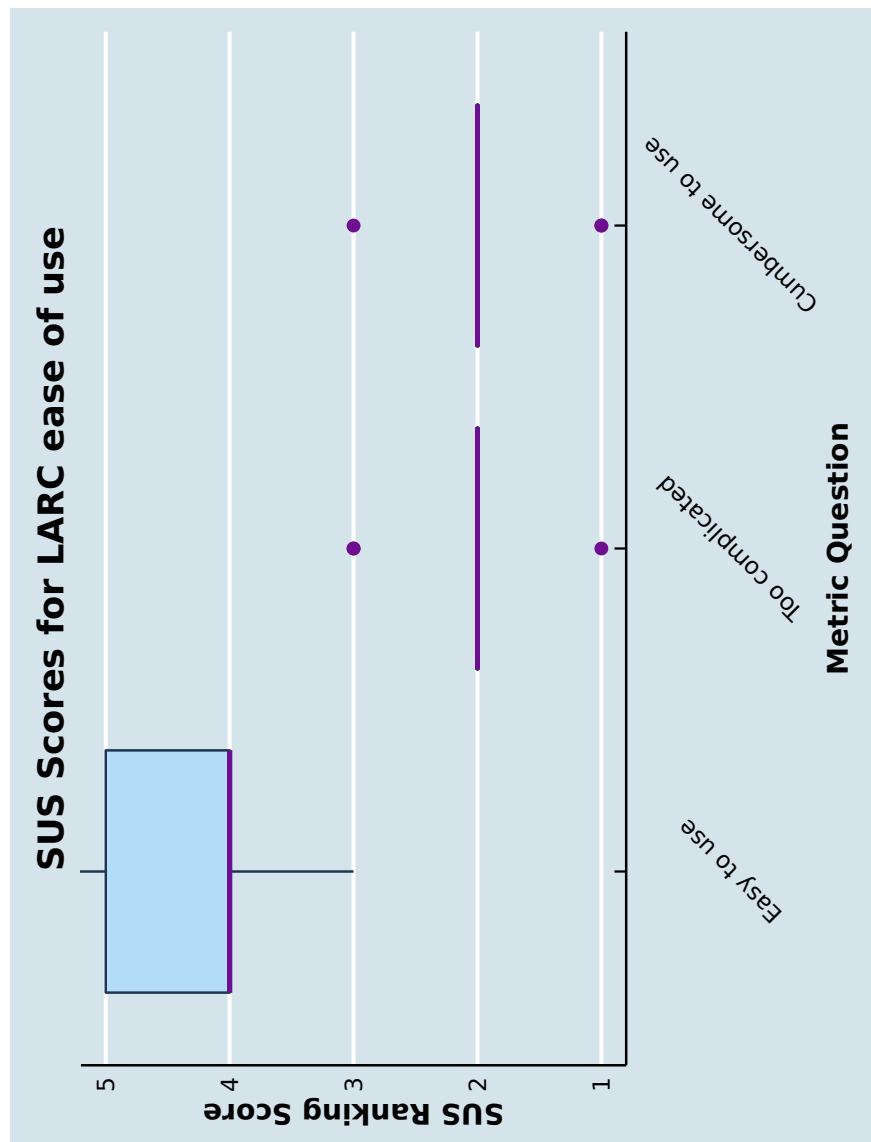


Figure 7.2: Scores from the evaluation participants on SUS questions relating to the ease of use of the LARC prototype. The scoring scheme is a five point Likert scale from 1 (strongly disagree) to 5 (strongly agree). See **Section 7.7.1** for an explanation of the diagram.

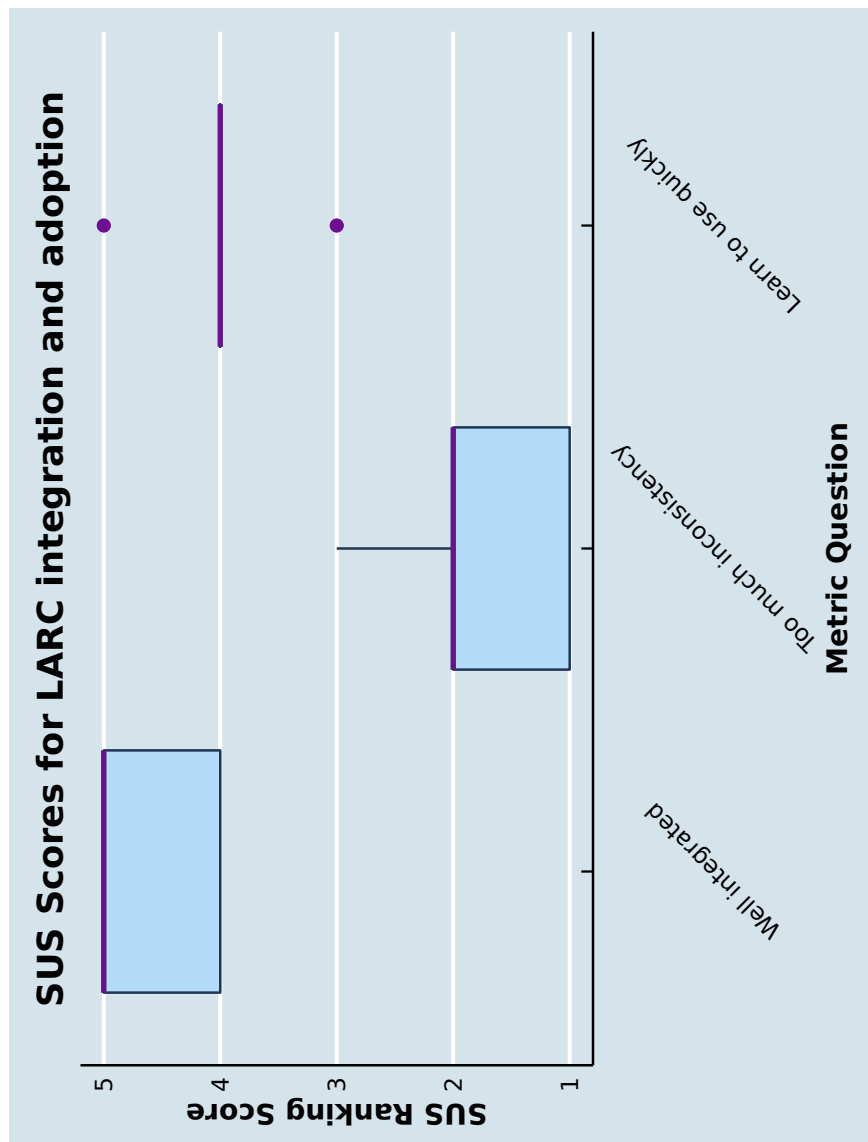


Figure 7.3: Scores from the evaluation participants on SUS questions relating to the quality of integration and ease of adoption of the LARC prototype. The scoring scheme is a five point Likert scale from 1 (strongly disagree) to 5 (strongly agree). See [Section 7.7.1](#) for an explanation of the diagram.

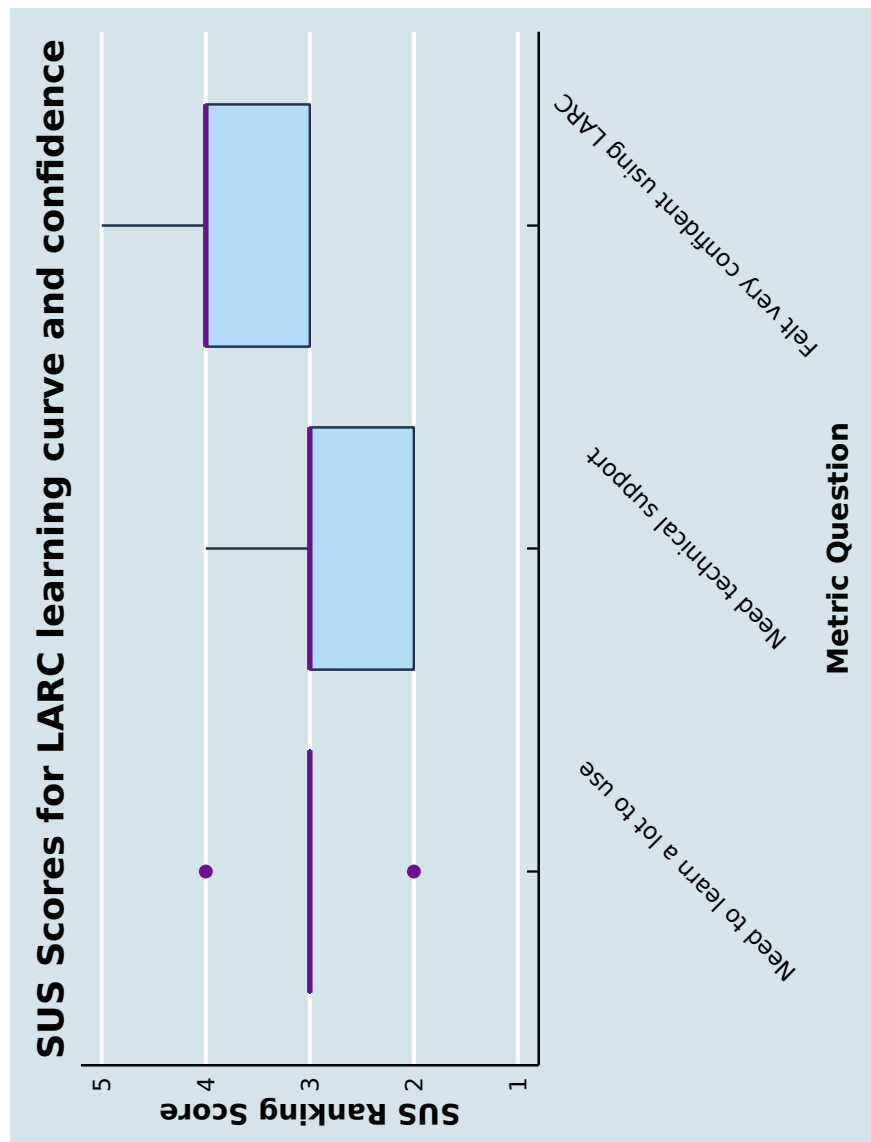


Figure 7.4: Scores from the evaluation participants on SUS questions relating to the learning curve of and user confidence in using the LARC prototype. The scoring scheme is a five point Likert scale from 1 (strongly disagree) to 5 (strongly agree). See **Section 7.7.1** for an explanation of the diagram.

7.7 Results

7.7.1 Quantitative metrics - SUS ratings for LARC

Scores for the effectiveness of the LARC system across the different metrics of the System Usability Scale are presented as box plots in Figure 7.1 (system utility), Figure 7.2 (ease of use), Figure 7.3 (effectiveness of integration in the software) and Figure 7.4 (user confidence in using the software). As noted in the figure captions, the SUS scale runs from 1 (strongly disagree) to 5 (strongly agree). The box plots show the median (bold line), 25th and 75th percentiles (quartiles) - the outer boxes, further interquartile ranges (IQR) within the data (whiskers) and maximum and minimum values (points) of the survey responses to each question. Where there was limited range in the survey responses (because most respondents picked the same response on the scale), some or all of these values may be identical. In these cases the values are given visual precedence in the order median-quartile-IQR-maximum/minimum. Hence some of the data points give rise just to maximum and minimum points, and/or a median line, rather than a box representation.

The overall System Usability Scores for the LARC system were as shown in Table 7.3, organised by participant.

Participant	Score
Participant 1	80
Participant 2	70
Participant 3	70
Participant 4	62.5
Participant 5	70
Participant 6	80
Participant 7	87.5
Participant 8	75
Participant 9	75
Average SUS score	74.4

Table 7.3: Overall SUS scores for the LARC system

The System Usability guidelines in [23] state that any system with an average usability factor of more than **68** displays above-average ease of use and system utility. The derived score for LARC indicates a **B** grade on the overall metric. It is interesting to note that the system was much more positively received by professional lawyers than it was by law students. The results show that

the strangeness of the language search interface in comparison to a Google-type approach to information retrieval contributed heavily to the lower overall usability scores. There is a correlation evident in the data between those who watched introductory videos for LARC which were prepared and published by the researcher in advance of the evaluation and higher overall usability ratings for the system.

7.7.2 Qualitative measures

This section includes responses to the evaluation survey which were collected in the unrestricted text entry fields of the feedback form. The answers which are summarised here were drawn from responses to the following questions: what did you like most about the LARC software?; what did you dislike most about the LARC software?; how would you change LARC to make it more useful in your work?; how useful is the language search facility in LARC?; and which features of LARC would be most useful for law students and trainee lawyers?

7.7.2.1 Favourite features

It becomes clear from the evaluation feedback that the most welcome features of the LARC platform amongst the cohort of responders tend to be simple functionality. **EP7** said that *"...[t]he shopping basket is a great idea. I often find it difficult to maintain a record of what I have looked at - what was important - so saving what I looked at before is a good idea."* **EP1** contributes that *"...[t]he search facility for a Judge may be very relevant to find out a pattern of how or why a judge makes his decisions to try to determine which way he might look at the current facts to be judged."* The search facets and their separation of different data entry points for information retrieval was generally well received. **EP5** stated that their favourite feature was *"...that the cases connected to statutes were available for you to see at the side of the screen."* **EP6** found the citation display to be their most useful feature. They said that *"...[t]he colour gradient is useful. The positive, neutral and negative elements reflected with green, grey and red is an excellent way to deal with the varying aspects of how a brain works. I used mind maps a lot in my study of the law so this particularly suits my [style of learning]."* **EP2** said that *"...[t]he way you explained the "interesting phrases" [in the introductory video] with your terminology of frequent collocations, words that broadly diagnose, word combinations and diagnosing overall content is very descriptive of the possibilities. This function might be one of the gems of LARC."* The latter comment

relates to the display of interesting phrases for a particular case which can be accessed from the citation diagram in *Research View*.

EP7 also liked the simple search facility within the text of a single case report. They said “...*the search facility within one case is a very useful feature. I wish this had been available when I was studying and training.*” The idea of placing a collaborative document editor at the heart of the LARC system was also well received by some respondents in particular. **EP3** stated that “[t]he document processor would be good to help speed up the design of a brief with a few people. The authorship colour is useful. It takes facilities available with PDF [and] applies them for a few people to use it at the same time.” There is evidence from the response data that the advanced functionality of the document editor is too hidden, or not emphasised sufficiently, in the interface, however. **EP4** found this to be a factor in particular, asking rhetorically “...[w]ith the document editor, I understand that you would need to keep the functions simple, but is there a facility to save the work to come back later? If so, what if one person has spent more time on the project than another? Does it reflect time spent or an allocation of time spent so as to reflect how much of a percentage of the brief was prepared by each of the parties involved?” However, **EP6** managed to find the statistics and timeline functionality in the interface. They said “...[t]he statistics view to see the contributors to the document, joining up the colours, who is contributing the most and who was driving the research is a good idea, as is the timeline view and the navigable tool. This is important for those that have to bill hours spent.”

EP9 liked the citation view and the overall flexibility of the case law search options in the LARC platform. “All the various possibilities to direct research are adjustable. You can search by cases or judges etc. Chart and case connections with colours are really impressive. I think the chat facility is a substantial idea as a tool to communicate with other parties who are researching within the same areas.”

EP9

7.7.2.2 Least useful features

Negative comments about the LARC system as a whole were relatively rare in the result set. Although the platform scores highly for the utility of its integrated approach in the System Usability Scale benchmarks (see Section 7.7.1), this was not universally appreciated. **EP5** stated that their least favourite aspect of LARC was that “[t]here is a lot on the screen whilst you are researching so it is difficult to concentrate on what you need to be looking at.” **EP7** said that “...if you were to think

about marketing the product you would definitely need to use more practical examples of how useful it is and not so many academic terms in the interface or introductory videos which focus on examples from linguistics.” EP1 expanded on the issue of terminology in the interface by saying “...using simpler language to get the meaning across so as not to intimidate [users] might be a good idea. Give a reason why they would want to identify “negligence” in a particular setting for example, this would help to explain all of the functions on offer showing a result at the end.” EP6 contributes that “ “keep a concordance”, “export a concordance” and “queue it for export” - why are these tools useful? I think maybe that this part [of LARC] is too technical.” EP7 expands criticism of the technical terminology used to the approach taken in the introductory videos which were created and distributed to the evaluation group prior to the exercise. They say “...[w]hen you talk about words being proximal, I would have thought that a “phrase” is a better way of looking at it. When you say the aim is the need to have more specificity in search results, I think this is the crux.”

Many of the drawbacks that were highlighted with the LARC prototype were technical in nature rather than issues about the general approach taken to legal research. EP2 noted that the system layout, in particular for *Research view*, was not best presented on a small laptop screen. They said that “*case titles get chopped off on my laptop and the screen generally tries to jam a lot into a small area.*” This relates to the fact that the layout engine in the software is responsive but that this is achieved in a “brute force” manner at present by curtailing content ranges rather than by modifying layouts to be more readable in small viewports. The solution here might be to segment the interface more effectively through the use of views and overlays which can be triggered contextually. EP3 said that “*I like the tumbler switching mechanism switching it to Acts, this way of accessing information only when you need it could be more consistently used.*” EP4 offers more context to the idea of adapting screen layout for different devices, stating that “*I think design-wise, if the Document to Edit Box and Chat Box were located on the same side e.g. the left hand side that would leave more room for displaying legal information e.g. a case citation, a case report and a statute. Though I appreciate space is limited.*” This shows that the reconfigurable interface positioning features may be too hidden or not well explained. Some of the existing design decisions would have to be addressed differently with the required layout changes here, however. EP6 said that “*...[t]he easier to read function is a life saver as the text in the little boxes is not very conducive to the overall research process especially when you are well versed in speed reading a text. Lawyers are proficient scanners of text. So popping the panel out is critical whilst still*

having the same search facility."

7.7.2.3 Language search

The centrality of the language search functionality in LARC was generally well received by the evaluation participants. **EP7** said that *"...legal language is fluid, interpretative, emotional, subjective, objective, leading, cautionary, bold, definable. It is not black and white especially when phrases are used. It cannot be termed a "dictionary" as the language of phrases is too fluid...For these reasons, I like the idea that understanding legal language is important in [the system]."* **EP1** notes that the iterative collocation and variation display *"...allows a mixture of a wide cross section of terms e.g. "family" and "child" and "guardianship" and "BIS II" and "maintenance" and "in camera"...to whittle down all of the different strands of law including all of these terms."* **EP6** echoed those sentiments, saying that *"you need a broad understanding of [judicial] reasoning in an area of law. I would need to become familiar with [a broad range of previous treatment] in order to give the best advice possible."* **EP2** said that *"I find [existing legal information systems] very daunting as there is too much content on the pages, it is difficult to navigate as only a student and the language is like a new science that I have to learn to be comfortable with by myself. I think that [LARC] would be a big help to feeling easier about working with the language [of the law]."*

The specifics of the language search implementation in LARC attracted both positive and negative comments from different respondents. This variability was particularly evident in submissions from the four students in the cohort. **EP3** said that *"...it seems crazy that there is a limit to the potential length of a search possibility."* This comment relates to the default length of concordance lines in language search, which display 80 characters to the left of the query node and 80 characters to the right. The use of the full co-text option might ameliorate this problem in practice but that facility is clearly not sufficiently evident in the interface. **EP4** said that *"the [collocation] drill down option is great, it is different and novel. The query history and the dead end aspect are good facilities and would need to be used to get rid of redundant searches, to help keep [my] brain clear from clutter apart from anything else."* They continued by saying that *"I find the language of "collocate" and "concordance" difficult to grasp. I came out of LARC and watched the [introductory video on language search] and now I think I understand it better."* **EP5** was less positive about the language search interface as a whole. Their short response to this question said *"I found this aspect quite confusing and I was unsure how to use it."* **EP4** said that *"I am afraid I was a bit lost with the options for the "Normal rank", "Typical rank" and "Typical with Filter*

rank”.”

EP8 said that *“[l]anguage search in LARC is a really great and useful function to research precedents by exploring with collocations. I know language is sometimes quite a challenge for [lawyers] and I even personally make up my own collocations which are not proper sometimes in the context of previous language. Because of that I may struggle with my research not producing relevant results. That is why I like the exploration facilities a lot.”*

In the case of this participant, part of the attraction of the language search facilities seems to be because English was not their first language, although they were qualified to practice English law. Thus the guided system of searching through identifying frequent and properly constructed collocations helped to bridge any uncertainty about appropriate phrasal construction which may have existed.

7.7.2.4 Change requests

One of the questions in the LARC evaluation survey requested information about how the system could be changed in order to be more useful in legal research work. Responses to this question were not received from all participants, unlike the other data points. However, those respondents who did reply here highlighted a range of different improvements which were more or less specific to their current job roles in the legal sector. **EP7** highlighted a need for the system to provide more information about the areas of law which different instruments relate to. This has been implemented to an extent in legislation search through the *subjects* search facet. However, they said that *“the main thing that a professional lawyer would need to see on a basic search engine would be “what kind of law is it?” or “what area of law does it deal with?” For example, criminal, civil, commercial, banking, intellectual property. A facility to search “inter-disciplines” would be useful making research quicker and more concise to what the researcher needs to get at.”* This feature could in fact be implemented through the data already available in LARC. The case content from the British and Irish Legal Information Institute is grouped loosely by law report series, which gives an idea of the responsibilities of the different tribunals and courts involved in delivering judgements.

EP1 suggested that, whilst the existing functionality for saving cases, statutes and items of linguistic interest in a shopping basket was useful, this function needs to be customisable. They said *“[t]he researcher will need to give it [the session] a memorable title, like the precise legal area, the name of the client, the subject matter. If they use it a lot for 20 different clients, pieces of research and so on.”* This ties in with

a request from **EP6**, who stated that practising lawyers would need to be able to marry up content in LARC with existing client files and billing records in existing external systems for client and case management. *“We use an established case management software tool and I would be concerned that case materials which were created in [LARC] should be accessible and interoperable with that software. I appreciate that you can export documents in different formats but some form of automatic [synchronisation] would be essential when this is made into a product.”* **EP7** reiterated the importance of compliance and interoperability in this area. They said *“[a]s with the Word and PDF I suppose linking this into a billable programme to save time on typing it up separately would be another arm to expand upon.”*

EP9 wanted the system to be extended to include template documents and starting points for common types of written submission. *“...you might consider in the future expanding LARC to add a base of the most common patterns of legal written statements (documents) of claims or defences in a court action. It would help to compose a proper letter/document/statement for new students/lawyers who are not yet familiar with a letter’s construction.”* **EP9** suggested that the interactive scale for selecting date ranges in the facet to display cases by date could be expanded to other sections of the system in order to make searching against other properties more flexible. *“...when you search cases by court or by judge there is no date tool. I think it might be useful to add it. I find it really useful to manipulate the scale when choosing years that I am interested in so it could also work with searching other sources.”*

7.7.3 Critical incidents

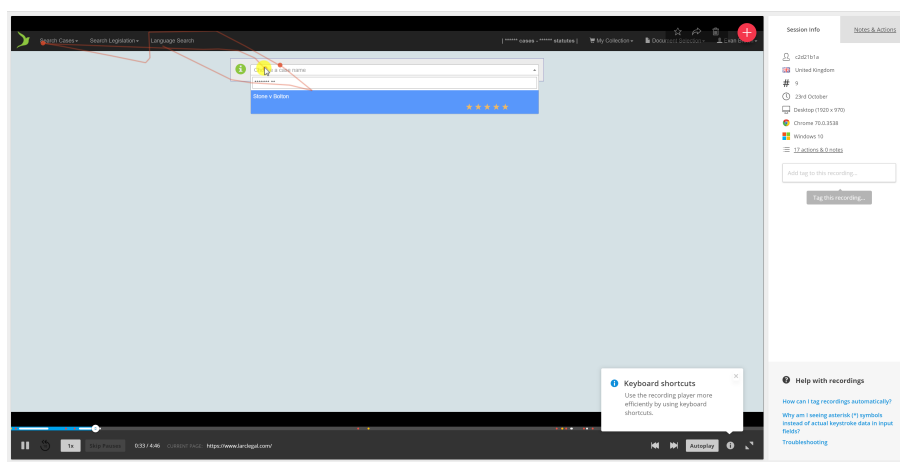


Figure 7.5: The evaluation session recording and playback interface.

In this section, the results of an analysis of the evaluation session video recordings is presented. The objective here is to identify areas of the LARC interface which seem to be unclear to users. Their behaviour when using different elements of the platform is evaluated and situations where there appears to be uncertainty or a loss of direction are highlighted as critical incidents. The results here come from annotation of the different video sessions that were stored during the evaluation period. See Figure 7.5 for a view of one recording playback session. Potential solutions to the critical incidents which have been identified are discussed in each case. In the following Section 7.7.4, these incident reports will be prioritised for remedial action according to their severity and repeatability.

Information icon lacks guidance: When users first log in to the LARC system, they are presented with a search facet screen to find an initial case by its title. This screen shares a basic design with the other search facets. The text input field sits to the right of a green information icon. Hovering over this icon displays explanatory text about how to conduct an initial search against a particular facet. A design decision was taken to make the text input field stand alone in this view. However, the recordings show that users moved their mice around the screen a lot at this stage without any apparent direction or purpose. It appears that the function of the stylised information icon is not immediately clear in many cases. The fix here would be either to place explanatory text directly in the page above the input field or to have a step-through presentation on first login which highlights different parts of the search interface in sequence and explains their function.

No way to select and load a document without first identifying a case: The evaluation videos show that many users lack a decisive research direction when they first login to LARC. The system has been designed such that an initial case or statute of interest must be chosen before the user is taken to *Research View*. However, it is possible that many users will not have a specific research direction, case or statute of interest in mind when they start using LARC. There needs to be an option to go straight to *Research view* and to load a document without selecting an accompanying case or statute first.

Interface not responsive enough on small viewports: The recording system which has been used to store videos of evaluation sessions replicates the layout of the screen in which it is embedded at the resolution that the user system implements. An issue which arises from these records is that the LARC interface does not respond to different viewport sizes as fluidly as would be desirable. A

potential fix here is to display different combinations of content on the screen for resolutions below a certain fidelity. The missing panes from the interface could then be made available to the user as modal dialogues.

Terminology in language search is too technical: The evaluation videos demonstrate that some users lose their direction in the language search interface. Most manage to implement an initial search query but there is little evidence of exploration after that point. A significant amount of indecisive mouse movement is also evidenced in some cases. In general, terminology in language search follows standards from corpus linguistics. These may be too technical and domain specific for non-expert users. The fix here is to devise a vocabulary for language search which is more easily understandable whilst maintaining the overall meaning of different choices in menus and in the interface as a whole.

Modal dialogue boxes too small and not responsive enough: As the relative lack of responsiveness in the LARC interface as a whole has previously been discussed, it is no surprise that the search facet views from *Research view* also need to be more fluid and adaptable within the interface. The user can launch new information requests from the research screen but these are displayed in modal dialogues in order to preserve the state of the rest of the environment. Evaluation videos show that users have to scroll around these dialogues a lot in order to manage information requests and to synthesise results. To fix the problem of responsiveness, a different layout is probably needed for the modal dialogues.

Timeline and saved context functionality hidden: The evaluation videos show relatively little exploration of the timeline and saved contexts facility in LARC. There is a correlation here between participants who did not watch the introductory videos and those who did not examine these features. Shared contexts were particularly under-utilised. This may be mainly because the system was not being used in a synchronous collaboration scenario and that each user was evaluating LARC in isolation. The fix here might be to move saved contexts out of the timeline altogether into a standalone system menu.

Shopping basket sessions hidden and not configurable: A number of evaluation participants accessed the session storage facility through both the shopping basket and the session history dialogue. They attempted to click repeatedly on session entries, presumably in order to try and name them or to otherwise work with the codified data. This issue of a lack of interactivity in the sessions manager was also

raised in the qualitative feedback from respondents. This is a relatively easy fix which entails a controller action for changing session identifiers in the database and an additional asynchronous update call in the sessions view.

Panel moving icons too close to other function buttons: Several participants encountered an apparent error state in *Research view* which was caused by inadvertently moving panels around on the screen. The grid layout is configurable by clicking on a *Move* icon in each panel header and then dragging and dropping the panel to a new location on screen. This reflects a design decision that the interface should be customisable for different preferences and requirements. The grid update action should be more constrained in order to prevent panel movement outside of the range of existing panel locations. Secondly, the button to move a panel in the header is too close to icons for other functions and should be segregated in order to prevent accidental invocation.

Citation diagram is a blocking resource: Several users entered a fault state after clicking on a linked case in the citation view. The evaluation videos show that they clicked on a case name and then the system apparently became unresponsive for a period of time. This was demonstrated through a variety of subsequent clicks and keyboard events which had no effect on the system. The problem here is that, although the initial population of the citation diagram for cases mentioned in a root case is asynchronous, the loading of cases cited in a lower level of the tree relies currently on a synchronous data call. The fix here is to implement an asynchronous call for tree updates which triggers a pull-down alert that the diagram is updating.

Interesting phrases functionality not explicit: Several evaluation session videos show users examining the contextual menu of content exploration options from the citation diagram. The respondents spent a significant amount of time looking at the *Interesting phrases* for various cases. However, the facility for conducting a whole corpus search or a session search by clicking on a phrase was little used. This issue can be fixed by the provision of better introductory materials that are designed for lawyers, by lawyers which explain the use of LARC.

Insert citations into document adds citation to the end of the text: The contextual menu for each node in the citation diagram features an option to insert a citation for the selected case into the current document. Several users clicked on a location within the document that they were working on and then used the

option in the menu to insert a citation. However, the system causes confusion here by placing the citation at the end of the document regardless of current cursor position. This can be fixed by making insertions into the XML data of the pad directly, where current cursor position is recorded and updated regularly.

Insert citation is not interactive: The evaluation videos show frequent use of the facilities in the citation layout to insert case citations into both the current document and the chat window. After users have navigated away from the cited case, there appears to be an expectation that clicking on the old citation will take them back to the relevant case report. However, case and statute citations in both the document and the chat window are not interactive. This issue is difficult to fix because XMPP clients tend to prohibit the inclusion of HTML content for security reasons. Any such data is parsed out of messages and only plain text is returned and archived.

Citation index in Case Connections panel header is unclear: The qualitative feedback from the evaluation survey and some evidence from the video records indicate that the placement of a citation index in the case connections panel header is confusing. Users tend to hover over this interface element for more information but none is forthcoming. The system actually shows the citation index for the currently-loaded case in the case report panel. However, this should be relocated or equipped with an information overlay in order to make it clearer.

Lack of keyboard shortcuts: Many of the evaluation session videos show users attempting to submit searches in the facet layouts by pressing a key (presumably RETURN or ENTER) rather than by clicking on the appropriate submit button with their mouse. The implementation of views in LARC currently does not support keyboard shortcuts at all. This can easily be fixed by appending form tags to the search layouts and it would provide much more natural functionality for the system as a whole.

7.7.4 Priorities for fixing incidents and future work

Table 7.4 shows a proposed hierarchy of importance for fixing the critical incidents that have been identified in LARC. In general, issues which cause errors that are difficult to recover from are ranked more highly than those which are aesthetic or less significant. The highest level is **1 - immediate fix**. The lowest priority level is **3 - desirable feature**. It is proposed that this list and the severity measures should be

used to inform immediate future development activities with the LARC platform now that the project is available as an open source resource. See the LARC source code repository at <https://bitbucket.org/evbuk1/lawspider/src/master/>.

Issue	Priority
No way to select and load a document without first identifying a case	1
Terminology in language search is too technical	1
Panel moving icons too close to other function buttons	1
Precedent diagram is a blocking resource	1
Insert citations into document adds citation to the end of the text	1
Citation index in Case Connections panel header is unclear	1
Lack of keyboard shortcuts	1
Interface not responsive enough on small viewports	2
Modal dialogue boxes too small and not responsive enough	2
Interesting phrases functionality not explicit	2
Information icon lacks guidance	3
Timeline and saved context functionality hidden	3
Shopping basket sessions hidden and not configurable	3
Insert citation into chat is not interactive	3

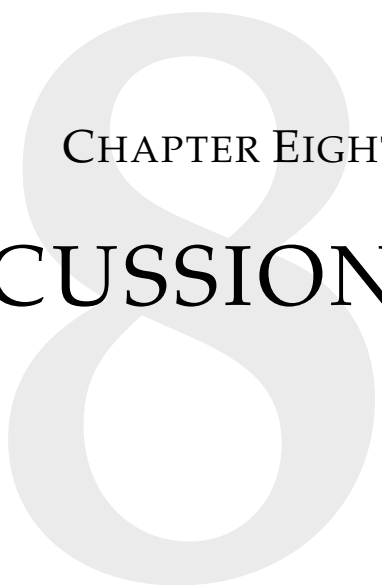
Table 7.4: Proposed fix priority of critical incident reports - on a scale of 1 - **immediate fix** to 3 - **desirable feature**

7.8 Conclusion

An evaluation of the Legal Research and Collaboration platform prototype has been presented in this chapter. The results are split between quantitative metrics which seek to establish the usability of the system and qualitative comments from the evaluation respondents. In the quantitative metrics, LARC scores highly in the evaluation group for their desire to use the system frequently. The software was generally not found to be overly complex and it fared well in terms of general ease of use. There is a caveat here in that many participants felt that they would need the assistance of a technical person in order to take full advantage of the different functionality that the platform provides. The close integration of system components was welcomed by the evaluators and this positive impression was reflected in the low scores for interface inconsistency. Most people felt that, with the right guidance, they could learn to use LARC quite quickly. The respondents did not find the software cumbersome but many did feel quite strongly that there was a lot to learn in aspects of the interface.

The qualitative comments that were received during the evaluation highlight that simple functionality in LARC accounted for the best-received features of the system. Integration was again praised and facilities like the search capability within case reports was found to be useful. The segregation and simple switching mechanism between dealing with case law and statute was well received. The central document editor emerged as a desirable feature particularly in the way in which it was tied in to other aspects of the system. Some of the advanced functionality for collaborative working, such as the statistics feature and the saved contexts browser, are apparently too deeply hidden from view for some respondents but others mentioned these features as useful. The biggest drawback for most users was the technical nature of words and phraseology associated with language search.

The critical incidents that have been identified from evaluation session video recordings feature a mixture of technical glitches and more fundamental issues with the system. Modifications like enabling users to access *Research view* and to load a document before selecting a case or statute of interest are important. There are also easier changes to make, however, including ensuring that the precedent diagram is asynchronous for loading on-demand nodes and tree structures.



CHAPTER EIGHT

DISCUSSION

8.1 Introduction

This chapter takes the results and other outcomes from the thesis as a whole and discusses them in relation to the main research question and process for answering it that has been outlined in previous chapters. A set of priorities for future work in creating open source collaboration tools for lawyers and methods through which corpus linguistics can and should be integrated into such systems is proposed. The entire LARC software package has been published as a publicly-accessible, open source project at <https://bitbucket.org/evbuk1/lawspider/src/master/>. The idea is that this chapter can form the basis of ongoing work to improve the software.

The high degree of connectivity between sources of legal information which exists predominantly as a result of the doctrine of precedent and the different ways that this facilitates processing and presenting that data to users is seen as a critical driver for future effort. The nature of this impetus for ongoing development activity within the open access data and open source software communities is considered. Finally, the potential success and barriers to adoption of LARC in these situations is evaluated with a discussion of the feedback that has been collected in the evaluation from Chapter 7.

8.2 Resolving the research question

The research question that is posed in this thesis is: **how can methodologies and approaches from corpus linguistics be integrated into a research platform for**

lawyers? This question has been addressed by following a formal investigative process which is proposed by Hartson and Pyla in [71]. In the context of this process, several procedural issues had to be addressed in order to tackle the overall question effectively.

Step in process	Purpose
P1	Consider how lawyers work in conducting legal research for case preparation
P2	Consider the role that technology plays in facilitating or hindering effective working practice
P3	Consider how open source software can be implemented as an integrated platform for legal research
P4	Consider how synchronous and asynchronous collaboration on legal documents can be supported by technology in an auditable manner
P5	Consider how search interfaces for linguistic content can be designed to target end-users who are not linguists or computer programmers
P6	Consider how usable the LARC software prototype is according to standardised user experience metrics
P7	Consider the feedback about LARC from lawyers and law students which should form the basis of future work on the platform

Table 8.1: A consolidated account of the process that has been followed in this thesis in order to answer the main research question.

Table 8.1 presents a consolidated list of the different steps in the user experience process which have been considered in the course of this thesis. The conclusions which have been drawn at each stage will be discussed in this section.

8.2.1 The results of P1 - How lawyers conduct legal research

This research suggests that lawyers work in teams very regularly as they conduct their work. The need for collaboration with others is less pronounced in university unless and until the student becomes involved in mooted activities. There are a large number of work roles inherent in presenting both real and moot legal cases. In university mooted, these roles were split on an ad-hoc basis between the two members of each team. In the profession, the same work roles exist although there tends to be more continuity in responsibilities as particular members of staff become expert in different tasks over the course of many cases. The most

significant difference in the skills that are required in practice when compared to training is an emphasis on project, client and discovery management.

All of the qualified solicitors in the respondent group in Chapter 4 highlighted the importance of organisational abilities of this type. The profession also involves collaboration between solicitors and barristers, and between the two sides in a legal matter - relationships which are often not simulated in mooting scenarios. Collaboration is driven mainly by email in practice whereas the law student may use instant messaging and other methods of real-time communication with a computer more in preparing a case. Collaboration in the profession is predominantly asynchronous and remote. Mooting teams often meet to prepare cases so their work is frequently asynchronous and colocated. There are relatively few scenarios in either training or practice where collaboration on work products is truly synchronous.

It is suggested that the nature and composition of legal work, especially at its nexus with information technology, is not well understood or satisfactorily explored at present. The work in this thesis makes a contribution to the current understanding. However, there is scope for a more detailed exposition of legal work. Current longitudinal studies are frequently historical and do not properly take account of the changes which computers have rendered in the legal environment. Others look to the future and consider the potential impact of developments like machine learning, pattern recognition and other forms of artificial intelligence. There seems to be less of an emphasis on understanding how widespread technological solutions have impacted upon the work of the lawyer and what benefits and problems these have created.

8.2.2 The results of P2 - Technology and working practice

Computerisation and the Internet has moved the search for and retrieval of legal information online. This new medium effectively makes traditional legal libraries of printed case reports and other sources obsolete. The coverage of information which has been achieved by existing platforms brings an unprecedented scope of sources to the individual legal research scenario. The development of legal information platforms prioritises content integration and coverage. This leads to large databases which are easy to query and which provide a guarantee of completeness that is unprecedented in the context of traditional research methods. However, technology also brings new barriers to productivity which are not well

answered by existing systems. Content integration leads to complex user interfaces which are difficult to use. Training in university tends to focus on a small range of features from the different platforms without properly exploring the best search strategies that are available to an individual so that they can make sense of the information which is most relevant to a particular research question. Content integration also leads directly to information overload, both for the lawyer and for the court.

The information publishers rely on establishing brand loyalty in university through low cost licensing and basic training which means that new entrants to the legal profession are conditioned to use and expect particular tools when they become qualified. In other respects, the lawyer is forced to use a jigsaw of different tools for document drafting, project management, case management, discovery, routine communication and all the other aspects of their modern working practice. The heterogeneous mix of different platforms leads to a problem that information becomes stuck in various silos and the lawyer then needs to spend time reconstituting work products in different environments as they proceed with their work.

It is important to develop an understanding here of which barriers to effective working practice are introduced by technology because of the way it works and which barriers are the result of poor training in the use of the available tools. There may be a pedagogical point to make in relation to the nature and standard of computing education that is available to the average lawyer as they progress through their career. Any system which seeks either to integrate content from multiple sources or to integrate multiple discrete tools under a unified user experience model will necessitate some level of learning curve. It is suggested that integration efforts will succeed best under a platform which has a low barrier to entry and a low total cost of ownership for both students and legal firms. An open source solution can address these issues.

8.2.3 The results of P3 - Integrating open source software

The idea that an open source software and open access development and licensing model can lead to diversification in the legal information sector is based on the principle that legal data should be freely accessible. This means not only that case reports and legislation be available for open access by the public but also that the raw product of the data itself be published and licensed appropriately

for integration into other products. This position is starting to gain traction in various countries. However, it is suggested that the uniquely coercive nature of the English doctrine of precedent makes specialist legal information products desirable here because there is a greater onus of allowing lawyers to understand the previous state of the law in England than there is elsewhere.

The task of implementing open source solutions for lawyers has two key aspects. First, underlying schemas for legal information in the form of platform- and component-agnostic communication standards need to be developed in order to allow data to pass between the different elements of an integrated legal information system. Secondly, it is important for open source development to be based upon best-in-class software. It is an inescapable feature of many open source projects that they are not well enough supported and they do not have a sufficiently active developer and user community to enable their inclusion in a production environment for mission critical work in the law. As it stands, LARC scores highly for user satisfaction with the degree and nature of integration that has been achieved. It seems that the principle of using an integrated environment for legal information retrieval and document drafting will be welcomed by many lawyers and this is a strength that can be built upon in future products.

8.2.4 The results of P4 - Supporting collaboration

LARC supports synchronous collaboration through the central role of the document editor which can be used by multiple people at the same time. A preponderance of effort in legal research is concentrated on producing documents of one type or another both for mooted scenarios and in legal practice. LARC also enables various auditing tools like the statistics display and the timeline. These facilities are designed both to encourage transparency in the collaborative process and to allow senior lawyers and teachers to audit group work effectively. The saved contexts feature allows coordination of group effort by providing the ability to step backwards in time as a document is composed in order to arrive at a conclusion about the direction, strengths and weaknesses of overall research directions and the contributions of individual participants.

The key to promoting synchronous collaboration in the legal domain appears to be that the tools to enable it should be simple and should not involve a significant learning curve for users. The lawyers and law students who participated in the contextual inquiry and other survey work in Chapter 4 uniformly preferred simple

solutions which enabled them to expend effort on preparing work products rather than on learning how to use software. Simple applications were also seen to be more reliable and less prone to errors than alternatives with more complex functionality. Finally, it is worth saying that a key barrier to the uptake of tools for synchronous collaboration is the problem that data in different environments becomes siloed too easily. LARC provides facilities for exporting documents and other content in both PDF and Microsoft Word formats to address this.

8.2.5 The results of P5 - Designing search interfaces for lawyers

The key contention in the design of the linguistic search interface in LARC is that information retrieval should be a workflow. Relevance, recall and saliency metrics are all improved where there is iteration in the interface which allows users to progressively focus inwards from general information to a more specific set of search requirements. The promotion of the collocation as the smallest indivisible unit of meaning is key here and drill down enables unlimited refinement of the information need for an individual task. Although the extraction and presentation of collocations is not a new idea in search, their derivation in real time in relation to a query is more novel. The role of the concordance and the presentational paradigms that this brings from corpus linguistics to first isolate a meaning of interest and then to filter and work with longer contexts around the query is a central part of the iterative workflow which is implemented in LARC.

Another important aspect of the language search approach from corpus linguistics is that a search scenario should never be treated as an ultimate view of the data or be permitted to pose a dead end after which a new interrogation is the only option available to a user. LARC implements a query history for language search which ensures that lawyers can easily navigate back to previous search requests and forward again once a productive direction has been isolated. The session history augments this functionality by enabling the sharing of data way-marks between groups of users working on different documents. This ties in with previous research which indicates the utility of shared search histories as a tool for increasing productivity and accuracy in legal research.

It is suggested that all of the algorithms for result ranking and filtration which are proposed in this thesis should only be used to guide the user. They are the ultimate arbiter of importance and relevance. A problem which is starting to be appreciated with artificial intelligence approaches to data analysis and presentation is that

the results are not sufficiently predictable or explainable from first principles that are simple enough for an average user to understand. Thus we now have the development of a number of research projects to augment machine learning approaches with explainable algorithms and metrics. LARC allows for the manual ordering, sorting and filtering of concordances based upon an approach called *ad-hoc classes*. This functionality sits alongside the *Typical* stratification and allows the user to define groups of language hits which they consider to be related to one another with software assistance. These groups can then either be retained in the result set or discarded.

8.2.6 The results of P6 and P7 - LARC's usability and feedback for future work

LARC scores reasonably well in the System Usability Scale metrics which were collected from the evaluation group and reported in Chapter 7. All of the respondents agreed that they would like to use the system regularly and that it would be useful in their work. The complexity of the language search interface led some to say that they would need the assistance of a technical person in order to use the system to its full potential. Interface and algorithmic vocabulary was the prime driver of lower scores in some metrics than would be ideal. However, an average SUS score of 74 places the system above average in terms of overall usability.

Qualitative feedback was most positive about simple presentational and organisational features of the platform, like the visual citation index, the search facility within individual case reports and the ability to toggle between acts cited in cases and other cited cases. The level of integration and the centrality and operation of the collaborative document editor were also praised by multiple respondents. On the other hand, although the power and novelty of the language search interface was praised by several evaluators, the accessibility of these features can be improved in future work.

8.2.7 Answering the main research question - How to apply corpus linguistics to legal research

Finally, it is necessary to address the main research question in this thesis - **how can methodologies and approaches from corpus linguistics be integrated into**

a research platform for lawyers?. The language search functionality in LARC was quite well received in both System Usability Scale rankings and qualitative feedback metrics from Chapter 7. Other platforms tend to focus on exposing the ordinary meaning of language for use in the courtroom by expert witnesses. LARC takes a different approach by allowing lawyers to explore linguistic constructions as an integral part of the workflow involved in preparing legal documents. This centres on a corpus of legal language from case law and legislation rather than upon a corpus of standard written English.

The experience of developing LARC shows that there is a tension between the aspiration to keep corpora as clean and free from derived information as possible and reconciling the unstructured text search capabilities with structured information for enabling search facets and other elements of system functionality. The conclusion drawn here is that corpora should always be clean in order to enable their use in as many different platforms and application scenarios as possible. However, some augmentation of the linguistic data is unavoidable in order to enable the use of a corpus in an integrated software application where specific information about cases and statutory sources is stored in a relational model.

The implementation of different algorithms for collocate and concordance ranking in LARC was less well received by the evaluation group. The idea of promoting collocations as fundamental units of meaning in an iterative search workflow was less contentious than might have been expected. However, the different options for evaluating saliency were felt to be too complicated and technical. This is as much a question of language in the system interface as it is a fundamental problem with exposing different ranking methodologies. The problem here can be partly ameliorated through a change in vocabulary and partly through the provision of better introductory materials which are designed for consumption by lawyers.

Overall, the central approach that was taken to applying corpus linguistics to legal information was intended to produce a guided search interface which allowed the user to progressively refine their information need by specifying more and more precise collocations to encapsulate what they were interested in. This worked well and was welcomed by the evaluation group. The idea that there should be no dead ends in the search process, supported by a query history, a shopping basket for interesting language and a key cases display in LARC, means that traditional approaches to interfaces for corpus linguistics can be applied in software for

people other than linguists as long as some of the specialised vocabulary from that domain is replaced with alternatives that express what a particular function is for in more transparent language.

8.3 Conclusion

The procedural answers that have been proposed in this thesis whilst addressing the key research question have been consolidated and considered here. As an initial prototype, LARC represents the first attempt to integrate corpus linguistics into an integrated system for legal research. Whilst other platforms allow linguists and academics to analyse legal language, the end users for these solutions are not practising lawyers and the tools do not fit into a coherent part of the workflow for legal research. The results of the evaluation score the system as more usable than average. There is good qualitative response from the majority of system evaluators. Future work would profitably focus on making the language search interface and saliency algorithms more approachable.



CHAPTER NINE

CONCLUSION

The system of common law in England grew up over centuries, based initially on devolved Anglo Saxon laws, before being systematised and centralised. The common law system gives rise to a key principle that the law should be fixed and determined so as to remove the threat of the arbitrary use of power. The doctrine of *stare decisis*, or precedent, also serves to structure and link cases linguistically. It means that like cases should be decided alike. Precedent is a coercive and strong doctrine under English law.

This emphasis on understanding precedent can and should be achieved by fostering a detailed understanding of legal language in the system user, who should remain the ultimate trained arbiter of relevance and importance. Legal language is technical and syntactically complex. It comprises specialised legal terms with meanings that are specific to jurisprudence and also overloaded terms from everyday language. Preferences for approaches to and methods for legal research are formed at university during the initial stages of legal training. New tools should therefore be designed specifically for use both by students and by practising lawyers.

The traditional model of legal education, which persists from the late nineteenth century, is doctrinalist and teacher-led. There is relatively little emphasis on fostering and developing broad social skills like teamwork, collaboration and computer literacy. Simulations of legal scenarios and environments have been practised for many years, mostly as optional extensions to standard legal curricula. This problem-based learning strategy is exemplified strongly in the traditional discipline of mooting. Efforts to apply technology here centre on broadening participation by removing the need for face-to-face interaction so that remotely-

situated students can be involved. Results have been mixed, however, because of the new burdens and practical limitations which the technology itself introduces.

However, computer-based products for legal information retrieval have been available for decades. They are now ubiquitous and are replacing printed law libraries at an accelerating rate. The move to online repositories of legal information has prioritised information coverage and completeness over usability and effective curation. Many commentators agree that the exponential growth in different sources of legal information and the challenges of considering them effectively are eroding legal research skills. This thesis has demonstrated that the application of techniques and design paradigms from the domain of corpus linguistics can lead to more effective information retrieval products for lawyers.

The development of legal corpora by interested parties is facilitated greatly by the recent availability of free and open source corpus creation and management tools. Activities to bring these tools into the legal domain have traditionally been limited by restricted access to sufficiently large collections of legal data. The open source model must rely upon an impetus for enabling and publishing open access legal information, particularly under the English system. Most corpus managers implement some form of the KeyWord In Context paradigm to present search results. This is effective for a trained linguist but tree-based alternatives may be easier for laypeople to understand. They promote knowledge about linguistic variation around search query nodes. The move towards different visualisations for corpus search results could sit well with the broader effort to make law more visual and easier to understand.

This thesis shows that lawyers must be able to collaborate with others and to work in groups effectively, both in training and in practice. The various reported studies demonstrate that lawyers and support personnel work collaboratively very often. They take on many different work tasks and roles during their activities. Some of the tasks, like externalisation and note-taking, are currently sub-optimal and are conducted in a simplistic manner so that individual productivity is not overtly compromised. Collaboration is usually asynchronous and remote in the profession or asynchronous and co-located in mooting and education. Technology does not support synchronised and co-located collaboration in the legal sector well at present.

The studies also confirm that new entrants to the legal profession are often found to

have deficient legal research skills by their senior colleagues. This is partly due to the transference of general Internet search techniques into legal information tools. The finding that software tools actively contribute to the skills deficit is a novel result. From the three studies and the triangulated set of results, it is recommended that novel software developments for use by lawyers should focus on using open data; on integration to support the whole case preparation workflow; on supporting the collaborative creation of textual documents; on streamlining note-taking and knowledge synthesis using a computer; on implementing new search techniques which prioritise the context of hits; and on flexibility and customisation of user interfaces in tools which are internally hosted.

On the basis of these guidelines, the LARC platform for integrated legal research and digital note-taking in groups has been presented and described. Nine key barriers to effective working practice that current tools in this sector introduce have been addressed. LARC has been focussed on enabling collaboration; reducing the problem of information overload; enabling audit of work by tutors and senior lawyers and creating a curated research environment by preserving the idea of precedent and case hierarchy. The system seeks to achieve these improvements through information visualisation and frequency-based, empirical linguistic analysis techniques which are automated. A targeted evaluation of key features of the platform shows that it performs well in terms of coverage and parity of presentation with manually-curated tools. Challenges for the future include improving content completeness and building a development community around the platform.

A key feature of the LARC platform is a linguistic search and result presentation facility. This has been proposed and implemented as a partial solution to the problems of information overload and poor search relevance which were highlighted in the user surveys. The approach taken here is based upon algorithms and interface design norms from the domain of corpus linguistics. The focus on linguistic context and development over time from the study of corpora is particularly relevant to precedent-based legal studies. The methodologies have been used to create a guided search mechanism which is based upon presenting collocations as a fundamental unit of meaning.

Particular emphasis has been placed upon creating corpus resources for English law in an environment where little alternative data is currently available in appropriate quantities or formats for the creation of corpora. The resulting

databases are also designed to be living resources. This means that they can be updated over time as new legal information is published without significant technical or practical overheads. An important contention made here is that statistical saliency measures operating on a large, representative legal corpus are more defensible, explainable and predictable than complex probability calculations which are often inherent in approaches to machine learning and other artificial intelligence efforts.

LARC seeks to use the corpus to build an iterative search experience which concentrates initially on refining information need based upon drilling down into interesting collocations that are presented for a root query. There is a query history implementation in the software which should mean that users never encounter unrecoverable dead ends in their search requests. It is possible to navigate forwards and backwards through different queries if the information returned at any stage is not interesting or irrelevant.

Concordances are presented by default in the traditional KeyWord In Context (KWIC) display format. This KWIC approach is not ideal, however, for looking at linguistic variation outside the node. Variation is important legally because it often denotes different applications and treatments of legal thought and principle. Sinclair's work suggests that a hierarchical visualisation can be used to better effect for this purpose. LARC therefore implements a variation on the *DoubleTree* layout which differs from the canonical version by preserving an idea of the sentence.

An evaluation of the LARC platform prototype shows that it scores highly as a system which lawyers want to use frequently. An overall score of 74 on the System Usability Scale places the software above average in terms of usability. The software was generally not found to be overly complex and it fared well in terms of general ease of use. The close integration of system components was welcomed by the evaluators and this positive impression was reflected in the low scores for interface inconsistency. Most people felt that, with the right guidance, they could learn to use LARC quite quickly.

Simple functionality accounted for the best-received features of the system. Facilities like the search capability within case reports was found to be useful. The segregation and simple switching mechanism between dealing with case law and statute was well received. The central document editor emerged as a desirable feature particularly because it was closely tied in to other aspects of the

system. The biggest drawback for most users was the technical nature of words and phraseology associated with language search. Whilst the potential of the approach taken here was evident to many in the group, it was felt that the alien terms that described different algorithms and features of the interface made for a sharp learning curve.

LARC represents the first attempt to include corpus linguistics in an integrated system for legal research. Whilst other platforms allow linguists and academics to analyse legal language, the end users for these solutions are not practising lawyers and the tools do not fit into a coherent part of the workflow for legal research. Future work would profitably focus on making the language search interface and terminology around saliency algorithms more intuitive. The critical incidents which were discovered and which created error states in the prototype encompass this work.

Future developments should preserve the tight and popular integration of the platform, should build on the advantage of simplicity by adding new features which are easy to use and should attempt to make the platform interoperable with existing solutions for legal case management and client relationship management. There is also a need to develop a portal for the platform which guides system use through introductory materials that are written for lawyers, by lawyers.

REFERENCES

- [1] Alex Aldridge. 2013. The Legal Education Training Review is finally here. And not much has changed. The Guardian. (2013). <https://www.theguardian.com/law/2013/jun/26/legal-education-training-review>
- [2] B. Ambrogi. 2018. New Corpus Linguistics Platform Lets Legal Researchers Explore the Meanings of Words and Phrases. (2018). <https://www.lawsitesblog.com/2018/09/new-corpus-linguistics-platform-lets-legal-researchers-explore-meanings-words-phrases.html>
- [3] Amicus ID. 2019. Amicus Resolution. (2019). <https://amicus.co/>
- [4] Laurence Anthony. 2019. AntConc. (2019). <https://www.laurenceanthony.net/software/antconc/>
- [5] Olufunmilayo B. Arewa. 2006. Open Access in a Closed Universe: Lexis, Westlaw, Law Schools, and the Legal Information Market. *Lewis & Clark Law Review* 10, 4 (2006), 797–836.
- [6] Aleks Aris and Ben Shneiderman. 2007. Designing semantic substrates for visual network exploration. *Information Visualization* 6, 4 (11 2007), 281–300.
- [7] W. H. G. Armytage. 1953. The conflict of ideas in English university education - 1850 - 1867. *Educational Theory* 3, 4 (1953), 327–343.
- [8] Anthony Baldrey. 2008. What are concordances for? Getting multimodal concordances to perform neat tricks in the university teaching and testing cycle. *EUT - Edizioni Universita di Trieste Research* (2008).
- [9] L. Bannon, M. Robinson, and K. Schmidt. 1991. Collaborative activity and technological design: Task coordination in London Underground control rooms. In *Proceedings of the Second European Conference on Computer-Supported Cooperative Work*. 65.

- [10] Steven M. Barkan, Barbara Bintliff, and Mary Whisner. 2015. *Fundamentals of legal research* (tenth ed.). Foundation Press.
- [11] Michael Barlow. 2019. ParaConc. (2019). <https://paraconc.com/>
- [12] Geoff Barnbrook, Oliver Mason, and Ramesh Krishnamurthy. 2013. *Collocation: Applications and implications*. Springer.
- [13] Gena R. Bennett. 2010. *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press.
- [14] David Berman. 2000. Toward a New Format for Canadian Legislation. *Human Resources Development Canada and Justice Canada Pilot Project* (2000). <https://www.davidberman.com/NewFormatForCanadianLegislation.pdf>
- [15] Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- [16] Tom Bingham. 2007. The rule of law. *Cambridge Law Journal* 66, March 2007 (2007), 67–85.
- [17] David C. Blair and Melvin E. Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM* 28, 3 (1985), 289–299.
- [18] Blog.dhananjaynene.com. 2008. Performance Comparison - C++/Java/Python/Ruby/Jython/JRuby/Groovy. (2008). <http://blog.dhananjaynene.com/2008/07/performance-comparison-c-java-python-ruby-jython-jruby-groovy/>
- [19] Jeanette Blomberg, Lucy Suchman, and Randall H. Trigg. 1996. Reflections on a work-oriented design project. *Human-Computer Interaction* 11, 3 (1996), 237–265.
- [20] John M. Bowers and Steven D. Benford. 1990. *Studies in Computer-Supported Cooperative Work: Theory, practice and design*. North-Holland Publishing Co.
- [21] L. Karl Branting. 2003. A reduction-graph model of precedent in legal analysis. *Artificial Intelligence* 150, 1-2 (11 2003), 59–95.
- [22] British Law Reference Corpus. 2018. BLaRC: British Law Reference Corpus. (2018). <https://www.sketchengine.eu/blarc-british-law-reference-corpus/>

-
- [23] John Brooke. 1996. SUS - A quick and dirty usability scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.
- [24] C. R. Brunschwig. 2014. On Visual Law: Visual Legal Communication Practices and their scholarly exploration. In *Zeichen und Zauber des Rechts: Festschrift für Friedrich Lachmayer, Erich Schweihofer et al.* (Ed.). Editions Weblaw. <http://www.rwi.uzh.ch/oe/zrf/abtrv/brunschwig/publications/ColetteRBrunschwigOnVisualLaw2014.pdf>
- [25] Paul D. Callister. 2003. Beyond training: Law librarianship’s quest for the pedagogy of legal research education. *Law Library Journal* 95 (2003), 7.
- [26] Caselaw Access Project. 2018. The Caselaw Access Project. (2018). <https://case.law/>
- [27] Casetext. 2017. Casetext. (2017). <https://casetext.com/>
- [28] H. E. Chandler. 2001. The Complexity of Online Groups: A Case Study of Asynchronous Collaboration. *ACM Journal of Computer Documentation* 25, 1 (2001), 17–24.
- [29] Robert P. Charrow and Veda R. Charrow. 1979. Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review* 79, 7 (1979), 1306–1374.
- [30] George C. Christie. 1963. Vagueness and legal language. *Minnesota Law Review* 48 (1963), 885.
- [31] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16, 1 (1990), 22–29.
- [32] James A. Coriden. 2004. *An introduction to canon law*. A&C Black.
- [33] Corpus-analysis.com. 2018. Tools for Corpus Linguistics. (2018). <https://corpus-analysis.com/>
- [34] Corpus of Historical English Law Reports (CHELR). 2018. Corpus of Historical English Law Reports. (2018). <http://www.helsinki.fi/varieng/CoRD/corpora/CHELAR/>
- [35] Cortext. 2019. CorpusExplorer. (2019). <https://docs.cortext.net/corpus-explorer/>

- [36] Malcolm Coulthard. 2005. The linguist as expert witness. *Linguistics and the Human Sciences* 1, 1 (2005), 39.
- [37] Fiona Cownie. 2003. Alternative values in legal education. *Legal Ethics* 6, 2 (2003), 159–174.
- [38] Michael Cross. 2018. LETR: law students not prepared for work, Society says. The Law Gazette. (2018). <https://www.lawgazette.co.uk/news/letr-law-students-not-prepared-for-work-society-says-/71505.article>
- [39] Rupert Cross and James William Harris. 1991. *Precedent in English law*. Clarendon Press.
- [40] Chris Culy and Verena Lyding. 2010a. Double Tree: An Advanced KWIC Visualization for Expert Users. *Fourteenth International Conference on Information Visualisation* (7 2010), 98–103.
- [41] Chris Culy and Verena Lyding. 2010b. Visualizations for exploratory corpus and text analysis. *Proceedings of the Second International Conference on Corpus Linguistics (CILC-10)* (2010), 257–268.
- [42] Chris Culy and Verena Lyding. 2011. Corpus Clouds - Facilitating text analysis by means of visualizations. *Human Language Technology. Challenges for Computer Science and Linguistics* (2011), 351–360.
- [43] Michael Curtotti and Eric McCreath. 2012. Enhancing the Visualization of Law. In *Law via the Internet Twentieth Anniversary Conference*, Cornell University.
- [44] Brenda Danet. 1980. Language in the legal process. *Law & Society Review* 14, 3 (1980), 445–564.
- [45] Brenda Danet. 1990. Language and law: An overview of 15 years of research. *Handbook of Language and Social Psychology* (1990), 537–560.
- [46] Mark Debenham. 2012. Mapping out legal research (JustCite). *Australian Law Librarian* 20 (2012), 153.
- [47] N. J. C. den Bergh. 1988. The interpretation of Statutes in Hermeneutical perspective. In *Law and Semiotics*. Springer, 341–357.

-
- [48] Albert Venn Dicey. 2013. *The Law of the Constitution*. Vol. 1. Oxford University Press.
- [49] Pierre Dillenbourg. 1999. What do you mean by collaborative learning? In *Collaborative learning: Cognitive and Computational Approaches*, P. Dillenbourg (Ed.). Elsevier, 1–19.
- [50] Alan Dix. 2018. Sufficient Reason. In *Workshop on HCD for Intelligent Environments, BHCI*. Computational Foundry, Swansea University.
- [51] Tanya Du Plessis and A. S. A. Du Toit. 2006. Knowledge management and legal practice. *International Journal of Information Management* 26, 5 (2006), 360–371.
- [52] Donald J. Dunn. 1993. Why legal research skills declined, or when two rights make a wrong. *Law Library Journal* 85 (1993), 49.
- [53] Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, 1 (1993), 61–74.
- [54] Mark Elliott and Robert Thomas. 2012. Tribunal justice and proportionate dispute resolution. *The Cambridge Law Journal* 71, 2 (2012), 297–324.
- [55] John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* (1957).
- [56] Brian Fitzgerald. 2006. The transformation of open source software. *MIS quarterly* (2006), 587–598.
- [57] Domenick L. Gabrielli. 1981. The importance of research and legal writing in the law school education. *Albany Law Review* 46 (1981), 1.
- [58] Jonas Gamalielsson and Björn Lundell. 2014. Sustainability of Open Source software communities beyond a fork: How and why has the LibreOffice project evolved? *Journal of Systems and Software* 89 (2014), 128–145.
- [59] Claire M. Germain. 2007. Legal Information Management in a Global and Digital Age: Revolution and Tradition. *Cornell Legal Studies Research Paper No. 07-005* (2007), 134–163.
- [60] John Gibbons. 1994. Language constructing law. In *Language and the Law*, John Gibbons (Ed.). Longman, 3–10.

- [61] John Gibbons. 1999. Language and the law. *Annual Review of Applied Linguistics* 19 (1999), 156–173.
- [62] H. Patrick Glenn. 1986. Persuasive authority. *McGill Law Journal* 32 (1986), 261.
- [63] Neal Goldfarb. 2017. A Lawyer’s Introduction to Meaning in the Framework of Corpus Linguistics. *Brigham Young University Law Review* (2017), 1359.
- [64] Rt Hon Lord Goldsmith. 2006. Government and the Rule of Law in the Modern Age. *Justice Journal* 3, 1 (2006).
- [65] Monica Goyal. 2017. Do Lawyers and Law Students Have the Technical Skills to Meet the Needs of Future Legal Jobs? (2017). <http://www.slaw.ca/2017/06/29/do-lawyers-and-law-students-have-the-technical-skills-to-meet-the-needs-of-future-legal-jobs/>
- [66] Stefan Th. Gries and Brian G. Slocum. 2017. Ordinary Meaning and Corpus Linguistics. *Brigham Young University Law Review* (2017), 1417.
- [67] Scott Griffith. 2015. Do Lawyers Distrust Technology? LinkedIn. (2015). <https://www.linkedin.com/pulse/do-lawyers-distrust-technology-scott-griffith/>
- [68] Joelle Grogan. 2018. Suffering from Withdrawal - Controversy in the UK EU (Withdrawal) Bill. (2018). <https://verfassungsblog.de/suffering-from-withdrawal-controversy-in-the-uk-eu-withdrawal-bill/>
- [69] Margaret Hagan. 2015. Open Law Lab: Visual Law. (2015). <http://www.openlawlab.com/project-topics/illustrated-law-visualizations/>
- [70] Helena Happio. 2017. Legaldesignjam. (2017). <http://legaldesignjam.com/>
- [71] Rex Hartson and Pardha S. Pyla. 2012. *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier.
- [72] F. A. Hayek. 2012. *Law, legislation and liberty: a new statement of the liberal principles of justice and political economy*. Routledge.
- [73] F. A. Hayek. 2014. *The Road to Serfdom*. Routledge.

- [74] Christian Heath and Paul Luff. 1992. Collaboration and control: Crisis management and multimedia technology in London Underground Line Control Rooms. *Computer Supported Cooperative Work (CSCW)* 1, 1-2 (1992), 69–94.
- [75] C. Heath, P. Luff, and M. Svensson. 2003. Technology and medical practice. *Sociology of Health & Illness* 25 (2003), 75–96.
- [76] Chris Heffer. 2002. "If you were Standing in Marks and Spencers": Narrativisation and Comprehension in the English Summing-up. In *Language in the Legal process*, J. Cotterill (Ed.). Springer, 228–245.
- [77] Jim Henderson, Evan Brown, and Chris Mitchell. 2005. Adapting Open Source Software for Education: Challenges, methodologies and results. *Open Source for Education in Europe* (2005), 169.
- [78] Carissa Byrne Hessick. 2017. Corpus Linguistics and the Criminal Law. *Brigham Young University Law Review* (2017), 1503.
- [79] Kenneth J. Hirsh and Wayne Miller. 2003. Law School Education in the 21st Century: Adding Information Technology Instruction to the Curriculum. *William & Mary Bill of Rights Journal* 12 (2003), 873.
- [80] HM Courts & Tribunals Service. 2010. Solicitors' guideline hourly rates. In *Crime Justice and Law Guidance*. UK Government. <https://www.gov.uk/guidance/solicitors-guideline-hourly-rates>.
- [81] Daniel Hoadley. 2018. Open access to case law - how do we get there? Infolaw. (2018). <http://www.infolaw.co.uk/newsletter/2018/11/open-access-case-law-get/>
- [82] Nick Holmes. 2012. Who owns copyright in judgments? *Infolaw Newsletter* (2012). <https://www.infolaw.co.uk/newsletter/2012/03/who-owns-copyright-in-judgments/>
- [83] James Clarke Holt, George Garnett, and John Hudson. 2015. *Magna carta*. Cambridge University Press.
- [84] M. Huber. 2018. The Old Bailey Proceedings Corpus, 1674-1834. (2018). <http://www.helsinki.fi/varieng/series/volumes/01/huber/>

- [85] John Hudson. 2014. *The formation of English common law: law and society in England from the Norman Conquest to Magna Carta*. Routledge.
- [86] ICLR. 2019. The Incorporated Council of Law Reporting. (2019). <https://www.iclr.co.uk/about/>
- [87] Yasu Imao. 2019. CasualConc. (2019). <https://sites.google.com/site/casualconc/>
- [88] Hiroshi Ishii, Minoru Kobayashi, and Jonathan Grudin. 1993. Integration of interpersonal space and shared workspace: ClearBoard design and experiments. *ACM Transactions on Information Systems (TOIS)* 11, 4 (1993), 349–375.
- [89] Yolanda P. Jones. 2016. Expansive Legal Research. *International Journal of Legal Information* 44, 3 (2016), 241–268.
- [90] Justis Publishing. 2019. JustisOne. (2019). <https://app.justis.com/>
- [91] D. M. Katz. 2017. Computational Legal Studies. (2017). <https://www.computationallegalstudies.com/>
- [92] D. M. Kennedy and T. Mighell. 2018. The Lawyer’s Guide to Collaboration Tools and Technologies: Smart Ways to Work Together. In *The Lawyer’s Guide to Collaboration Tools and Technologies*. American Bar Association.
- [93] Mary Keyes and Richard Johnstone. 2004. Changing legal education: rhetoric, reality, and prospects for the future. *Sydney Law Review* 26, 4 (2004), 537.
- [94] Adam Kilgarrieff, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. ITRI-04-08: The Sketch Engine. *Information Technology* 105 (2004), 116.
- [95] Adam Kilgarrieff and David Tugwell. 2001. Word sketch: Extraction and display of significant collocations for lexicography. *Proceedings of the Collocations Workshop, ACL 2001* (2001), 32–38.
- [96] Jack Knight and Lee Epstein. 1996. The Norm of Stare Decisis. *American Journal of Political Science* 40, 4 (1996), 1018–1035.
- [97] Knomos. 2017. Knomos. (2017). <http://knomos.law/>

- [98] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate Detection using Shallow Text Features. *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (2010), 441–450.
- [99] Kohovolit.eu. 2017. Kohovolit: Analyses. (2017). <http://kohovolit.eu/en/category/analyses/>
- [100] Anita Komlodi and Wayne G. Lutters. 2008. Collaborative use of individual search histories. *Interacting with Computers* 20, 1 (2008), 184–198.
- [101] Anita Komlodi and Dagobert Soergel. 2002. Attorneys interacting with legal information systems: Tools for mental model building and task integration. *Proceedings of the American Society for Information Science and Technology* 39, 1 (2002), 152–163.
- [102] Steinar Kristoffersen and Fredrik Ljungberg. 1999. An empirical study of how people establish interaction: implications for CSCW session management models. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 1–8.
- [103] Herbert M. Kritzer. 1998. *Legal advocacy: Lawyers and nonlawyers at work*. University of Michigan Press.
- [104] C. C. Kuhlthau and S. L. Tama. 2001. Information search process of lawyers: a call for "just for me" information services. *Journal of Documentation* 57, 1 (2001), 25–43.
- [105] Miroslav Kurdov. 2017. SketchLex - Infographies Juridiques. (2017). <http://sketchlex.com/>
- [106] La Fabrique de la Loi. 2017. La Fabrique de la Loi. (2017). <https://www.lafabriquedelaloi.fr/>
- [107] Lancaster University. 2019. CQPWeb: Corpus Query Processor. (2019). <https://cqpweb.lancs.ac.uk/>
- [108] The Law Society. 2017. Yours, Mine or Ours? Effective collaboration in the legal industry. InsideOut Magazine. (July 2017). <http://communities.lawsociety.org.uk/in-house/insideout-magazine/july-2017/yours-mine-or-ours-four-steps-to-effective-collaboration/5062397.article>

- [109] Marlene J. Le Brun and Richard Johnstone. 1994. *The quiet (R)evolution: Improving student learning in law*. Law Book Co of Australasia.
- [110] Thomas R. Lee and Stephen C. Mouritsen. 2018. Judging Ordinary Meaning. *Yale Law Journal* 127, 4 (2018), 788–1105.
- [111] Philip Leith and Cynthia Fellows. 2013. BAILII, Legal Education and Open Access to Law. *European Journal of Law and Technology* 4, 1 (2013).
- [112] Lawrence Lessig. 2002. *The future of ideas: The fate of the commons in a connected world*. Vintage.
- [113] Nicola Lettieri, Antonio Altamura, and Delfina Malandrino. 2017. The legal macroscope: Experimenting with visual legal analytics. *Information Visualization* 16, 4 (2017), 332–345.
- [114] Lexical Analysis Software. 2019. WordSmith. (2019). <https://www.lexically.net/wordsmith/>
- [115] LexisNexis. 2019. LexisNexis Library UK. (2019). <https://internationalsales.lexisnexis.com/products/lexis-library>
- [116] House of Commons Library. 2017. Court Statistics for England and Wales. (2017). <http://researchbriefings.files.parliament.uk/documents/CBP-8372/CBP-8372.pdf>
- [117] Thomas Lundmark. 2012. The Methodology of Using Precedents. *University of Hull Research Repository* 9, 2012 (2012), 335–353.
- [118] Andrew Lynch. 1996. Mooting in Legal Education. *Legal Education Review* 7 (1996), 67–96.
- [119] Michael J. Lynch. 1997. An Impossible Task but Everybody Has To Do It—Teaching Legal Research in Law Schools. *Law Library Journal* 89 (1997), 415.
- [120] Paul Maharg. 2001. Negotiating the Web: Legal Skills Learning in a Virtual Community. *International Review of Law and Computers* 15, 3 (2001), 345–360.
- [121] Paul Maharg and Emma Nicol. 2014. Simulation and technology in legal education: a systematic review and future research programme. In *Legal education: Simulation in theory and practice*, C. Strevens, R. Grimes, and E. Phillips (Eds.). Ashgate.

-
- [122] Paul Maharg and Martin Owen. 2007. Simulations, learning and the metaverse: changing cultures in legal education. *Journal of Information, Law, Technology* 1 (2007).
- [123] Stephann Makri, Ann Blandford, and Anna L. Cox. 2008. Investigating the information-seeking behaviour of academic lawyers: From Ellis's model to design. *Information Processing & Management* 44, 2 (2008), 613–634.
- [124] Chris Marsden. 2019. Open Access to Law - How Soon? (2019). <https://www.scl.org/articles/3305-open-access-to-law-how-soon>
- [125] Catherine C. Marshall, Morgan N. Price, Gene Golovchinsky, and Bill N. Schilit. 2001. Designing e-books for legal research. In *Proceedings of the First ACM/IEEE-CS joint conference on Digital Libraries*. ACM, 41–48.
- [126] James Marson, Adam Wilson, and Mark Van Hoorebeek. 2005. The necessity of clinical legal education in university law schools: a UK Perspective. *Int'l J. Clinical Legal Educ.* 7 (2005), 29.
- [127] Susan Nevelow Mart. 2013. A Study of Attorneys' Legal Research Practices and Opinions of New Associates' Research Skills. *Task Force on Identifying Skills and Knowledge for Legal Practice* June (2013).
- [128] Masaryk University. 2019. NoSketch Engine. (2019). <https://nlp.fi.muni.cz/trac/noske>
- [129] Rowena Mason. 2018. EU withdrawal bill needs major rewrites, Lords committee says. (2018). <https://www.theguardian.com/politics/2018/jan/29/eu-withdrawal-bill-major-rewrites-house-of-lords-committee-brexite>
- [130] K Tamsin Maxwell and Burkhard Schafer. 2008. Concept and Context in Legal Information Retrieval. In *JURIX*. 63–72.
- [131] Douglas W. Maynard. 1983. Language in the Court. *Law & Social Inquiry* 8, 1 (1983), 211–222.
- [132] David Mellinkoff. 2004. *The language of the law*. Wipf and Stock Publishers.
- [133] Elizabeth Mertz. 1994. Legal language: Pragmatics, poetics, and social power. *Annual Review of Anthropology* 23, 1 (1994), 435–455.

- [134] Steve Mishkin. 2017. How Can Law Librarians Most Effectively Provide Legal Research Training? *Legal Information Management* 17, 1 (2017), 34–68.
- [135] Yasmin Morais. 2009. Scottish Legal History: A Research Guide. *NYULaw-Global* (2009).
- [136] Caleb Nelson. 2005. What is textualism? *Virginia Law Review* 91 (2005), 347.
- [137] John Henry Newman. 1859. *The scope and nature of university education*. Longman, Green, Longman, and Roberts.
- [138] Jacob Nielsen. 2000. Why You Only Need to Test with 5 Users. (2000). <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/>
- [139] Ikujiro Nonaka, Ryoko Toyama, and Noboru Konno. 2000. SECI, Ba and leadership: a unified model of dynamic knowledge creation. *Long range planning* 33, 1 (2000), 5–34.
- [140] Kees van Noortwijk. 2017. Integrated legal information retrieval: new developments and educational challenges. *European Journal of Law and Technology* 8, 1 (2017), 1–18.
- [141] Jean O’Grady. 2017. LexisNexis Acquires Ravel Law: A Tipping Point for Legal Analytics and the Second Wave of Legal KM. (2017). <https://www.deweybstrategic.com/2017/06/lexisnexis-acquires-ravel-law-tipping.html>
- [142] Open Data Commons. 2019a. ODC Open Database License (ODbL) Summary. (2019). <https://opendatacommons.org/licenses/odbl/summary/>
- [143] Open Data Commons. 2019b. ODC Public Domain Dedication and License Summary. (2019). <https://opendatacommons.org/licenses/pddl/summary/>
- [144] K. Pala, P. Rychly, and P. Smerk. 2010. Automatic Identification of Legal Terms in Czech Law Texts. In *Semantic Processing of Legal Texts*, E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia (Eds.). Springer, 83–94.
- [145] Abdul Paliwala. 2007. Legal e-Learning in Network Society. *Journal of Information Law and Technology* 1 (2007).

-
- [146] Dennis M. Patterson. 1988. Wittgenstein and the code: A theory of good faith performance and enforcement under article nine. *University of Pennsylvania Law Review* 137, 2 (1988), 335–429.
- [147] James C. Phillips and Jesse Egbert. 2017. Advancing Law and Corpus Linguistics: Importing Principles and Practices from Survey and Content-Analysis Methodologies to Improve Corpus Design and Analysis. *Brigham Young University Law Review* (2017), 1589.
- [148] Theodore Frank Thomas Plucknett. 2001. *A concise history of the common law*. The Lawbook Exchange, Ltd.
- [149] Jan Pomikálek, Miloš Jakubí, and Pavel Rychlý. 2012. Building a 70 billion word corpus of English from ClueWeb. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*. 502–506.
- [150] Roscoe Pound. 1907. Common law and legislation. *Harvard Law Review* 21 (1907), 383.
- [151] QLTS. 2016. Why English Law Governs Most International Commercial Contracts. (2016). <https://www.qlts.com/blog/why-english-law-governs-most-international-commercial-contracts>
- [152] Ravel Law. 2017. Ravel Law. (2017). <http://ravellaw.com/>
- [153] Joseph Raz. 2017. The rule of law and its virtue. In *The Rule of Law and the Separation of Powers*, R. Bellamy (Ed.). Routledge, 77–94.
- [154] Madhu C. Reddy, Paul Dourish, and Wanda Pratt. 2001. Coordinating Heterogeneous Work: Information and Representation in Medical Care. *Proceedings of European Conference on Computer Supported Cooperative Work*. (2001), 239–258.
- [155] John P. Rickards and Frank Friedman. 1978. The encoding versus the external storage hypothesis in note taking. *Contemporary Educational Psychology* 3, 2 (1978), 136 – 143.
- [156] P. Riehmann, H. Gruendl, M. Potthast, M. Trenkmann, B. Stein, and B. Froehlich. 2012. WORDGRAPH: Keyword-in-Context Visualization for NETSPEAK’s Wildcard Search. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1411–1423.

- [157] Pavel Rychlý. 2007. Manatee / Bonito - A Modular Corpus Manager.. In *RASLAN*. 65–70.
- [158] Milad Saad and Mary Lou Maher. 1996. Shared understanding in computer-supported collaborative design. *Computer-Aided Design* 28, 3 (1996), 183–192.
- [159] Andrew Sanders. 2015. Poor thinking, poor outcome? The future of the law degree after the Legal Education and Training Review and the case for socio-legalism. In *The futures of legal education and the legal profession*, Hilary Sommerlad, Sonia Harris-Short, Steven Vaughan, and Richard Young (Eds.). Hart, 139–169.
- [160] Judith N. Shklar. 1987. The Rule of Law: Ideal or Ideology. In *Political Theory and the Rule of Law*, Alan Hutchinson and Patrick Monahan (Eds.). Vol. 1. Carswell.
- [161] Marjorie A. Silver, Sanford Portnoy, and Jean Koh Peters. 2004. Stress, burnout, vicarious trauma, and other emotional realities in the lawyer-client relationship (Symposium: Lawyering and its discontents - Reclaiming meaning in the practice of law). *Touro Law Review* 19 (2004), 847–873.
- [162] John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press.
- [163] John Sinclair. 2004a. *How to use corpora in language teaching*. John Benjamins Publishing.
- [164] John Sinclair. 2004b. Trust the text. In *Trust the text*. Routledge, 19–33.
- [165] John Sinclair, Susan Jones, and Robert Daley. 2004. *English collocation studies: The OSTI report*. Bloomsbury Publishing.
- [166] Lawrence M. Solan and Tammy Gales. 2017. Corpus Linguistics as a Tool in Legal Interpretation. *Brigham Young University Law Review* (2017), 1311.
- [167] Robert W Solomon. 2009. Free and open source software for the manipulation of digital images. *American Journal of Roentgenology* 192, 6 (2009), W330–W334.
- [168] James A. Sprowl. 1976. Computer-Assisted Legal Research: Westlaw and Lexis. *Journal of the American Bar Association* 62, 3 (1976), 320–323.

- [169] Caroline Strevens and Roger Welch. 2014. Simulation and the learning of the law: constructing and using an online transactional assessment in employment law. In *Legal education: Simulation in theory and practice*, C. Strevens, R. Grimes, and E. Phillips (Eds.). Ashgate Publishing Limited, 43–66.
- [170] Gail Stygall. 2002. Textual barriers to United States immigration. In *Language in the Legal process*, J. Cotterill (Ed.). Springer, 35–53.
- [171] Richard E. Susskind. 2017. *Tomorrow's lawyers: An introduction to your future*. Oxford University Press.
- [172] Sweet and Maxwell. 2010. Blair: 54% more new laws every year than Thatcher. (2010). <https://www.sweetandmaxwell.co.uk/about-us/press-releases/260607.pdf>
- [173] Joshua Tauberer. 2018. How I changed the law with a GitHub pull request. Ars Technica. (2018). <https://arstechnica.com/tech-policy/2018/11/how-i-changed-the-law-with-a-github-pull-request/>
- [174] The British National Corpus. 2019. BNCWeb. (2019). <http://corpora.lancs.ac.uk/BNCweb/>
- [175] The Economic and Social Research Council. 1998. The Brown Corpus. (1998). <http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus>
- [176] The Open Knowledge Foundation. 2019. Open Definition 2.1 - Defining Open in Open Data, Open Content and Open Knowledge. (2019). <http://opendefinition.org/od/2.1/en/>
- [177] The Plain English Campaign. 2018. Legal Texts. (2018). <http://www.plainenglish.co.uk/campaigning/past-campaigns/legal.html>
- [178] The Spectator Archive. 1846. Legal Education. The Spectator. (1846). <http://archive.spectator.co.uk/article/24th-january-1846/13/legal-education>
- [179] The UK in a Changing Europe. 2017. Does EU law take precedence over UK law? (2017). <https://ukandeu.ac.uk/fact-figures/does-eu-law-take-precedence-over-uk-law/>

- [180] The UK Parliament. 1998. Human Rights Act 1998. (1998). <https://www.legislation.gov.uk/ukpga/1998/42>
- [181] Keir Thomas. 2006. *Beginning Ubuntu linux: From novice to professional*. Apress.
- [182] Thomson Reuters. 2019. Westlaw UK. (2019). <https://legalsolutions.thomsonreuters.co.uk/en/products-services/westlaw-uk.html>
- [183] Sarah Valentine. 2009. Legal research as a fundamental skill: A lifeboat for students and law schools. *The University of Baltimore Law Review* 39 (2009), 173.
- [184] Jaques Verrier. 2017. Lexmex.fr. (2017). <http://www.lexmex.fr/>
- [185] Friedemann Vogel, Hanjo Hamann, and Isabelle Gauer. 2018. Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies. *Law & Social Inquiry* 43, 4 (2018).
- [186] Martin Wattenberg and Fernanda B. Viégas. 2008. The Word Tree: an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1221–1228.
- [187] Louise Wayham. 2017. Open access and entrepreneurial reuse of UK legislation and case law in legal education technology. (2017). <https://blogdroiteuropeen.com/2017/11/03/open-access-and-entrepreneurial-reuse-of-uk-legislation-and-case-law-in-legal-education-technology-by-louise-wayham/>
- [188] Julian Webb. 2015. Response - A Tale of Two Cities: Reflecting on Lord Neuberger’s ‘Reforming Legal Education’. In *Perspectives on Legal Education*, N. Duncan, J. Guth, and C. Ashford (Eds.). Routledge, 38–56.
- [189] Wikipedia. 2017. List of landmark United Kingdom House of Lords cases. Wikipedia. (2017). https://en.wikipedia.org/wiki/List_of_landmark_United_Kingdom_House_of_Lords_cases
- [190] Margaret Ann Wilkinson. 2001. Information sources used by lawyers in problem solving: An empirical exploration. *Library & Information Science Research* 23, 3 (2001), 257–276.

- [191] Glanville Llewelyn Williams and A. T. H. Smith. 1982. *Learning the law*. HeinOnline.
- [192] Radboud Winkels. 1992. *Explorations in intelligent tutoring and help*. Vol. 15. IOS Press.
- [193] Radboud Winkels and others. 2015. The OpenLaws project: Big open legal data. In *Proceedings of the International Legal Informatics Symposium (IRIS 2015)*. 189–196.
- [194] Ludwig Wittgenstein. 2009. *Philosophical investigations*. John Wiley & Sons.
- [195] Jennifer Yule, Judith McNamara, and Mark Thomas. 2010. Mooting and Technology: To What Extent Does Using Technology Improve the Mooting Experience for Students? *Legal Education Review* 20, 1/2 (2010), 137–155.
- [196] Alina-Maria Zaharia. 2011. Noun phrases in legal language: an analysis of the problems posed by the translation of English multiple pre-modifying groups into Romanian. *Analele University Research Articles* (2011).
- [197] Staci Zaretski. 2012. Lawyers Sue Westlaw, Lexis-Nexis for Copyright Infringement. (2012). <https://abovethelaw.com/2012/02/lawyers-sue-westlaw-lexis-nexis-for-copyright-infringement/>



PART V

APPENDICES



APPENDIX A

SETTING THE SCENE: CONTEXTUAL INQUIRY, INTERVIEWS AND SURVEY

A.1 Mooting Work Roles

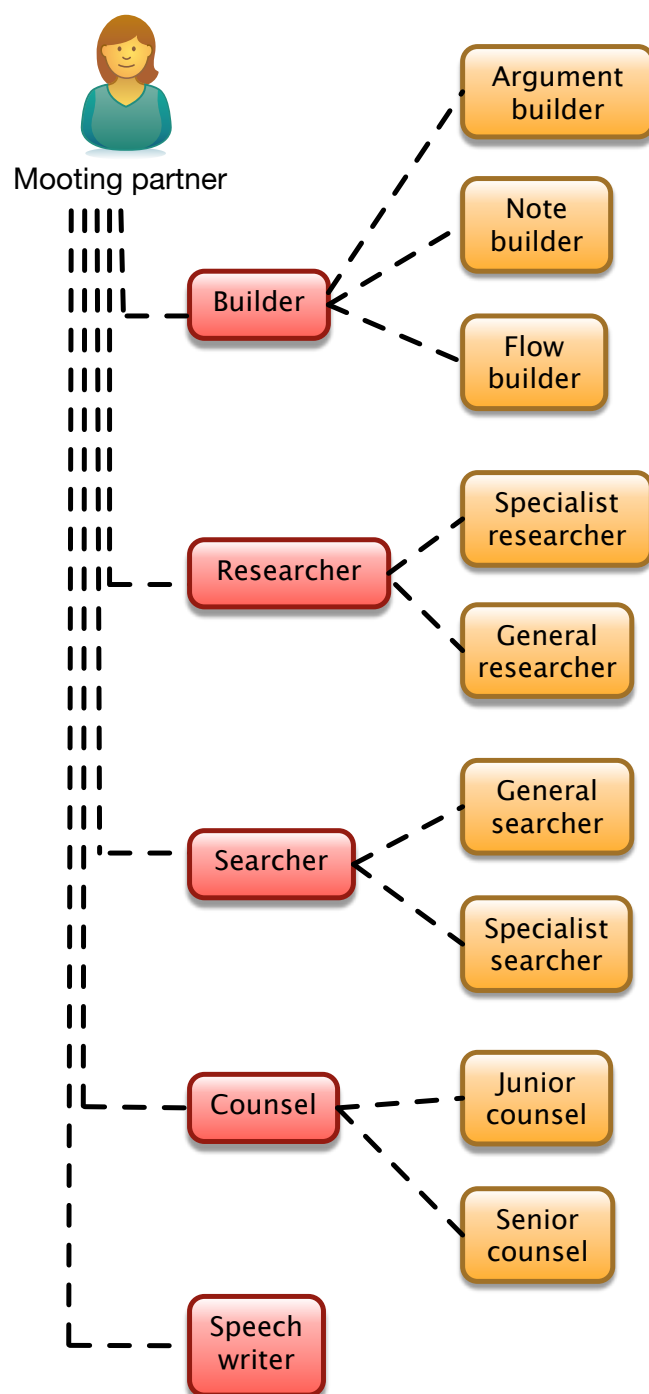


Figure A1: Work roles identified from video analysis of the mooting exercises

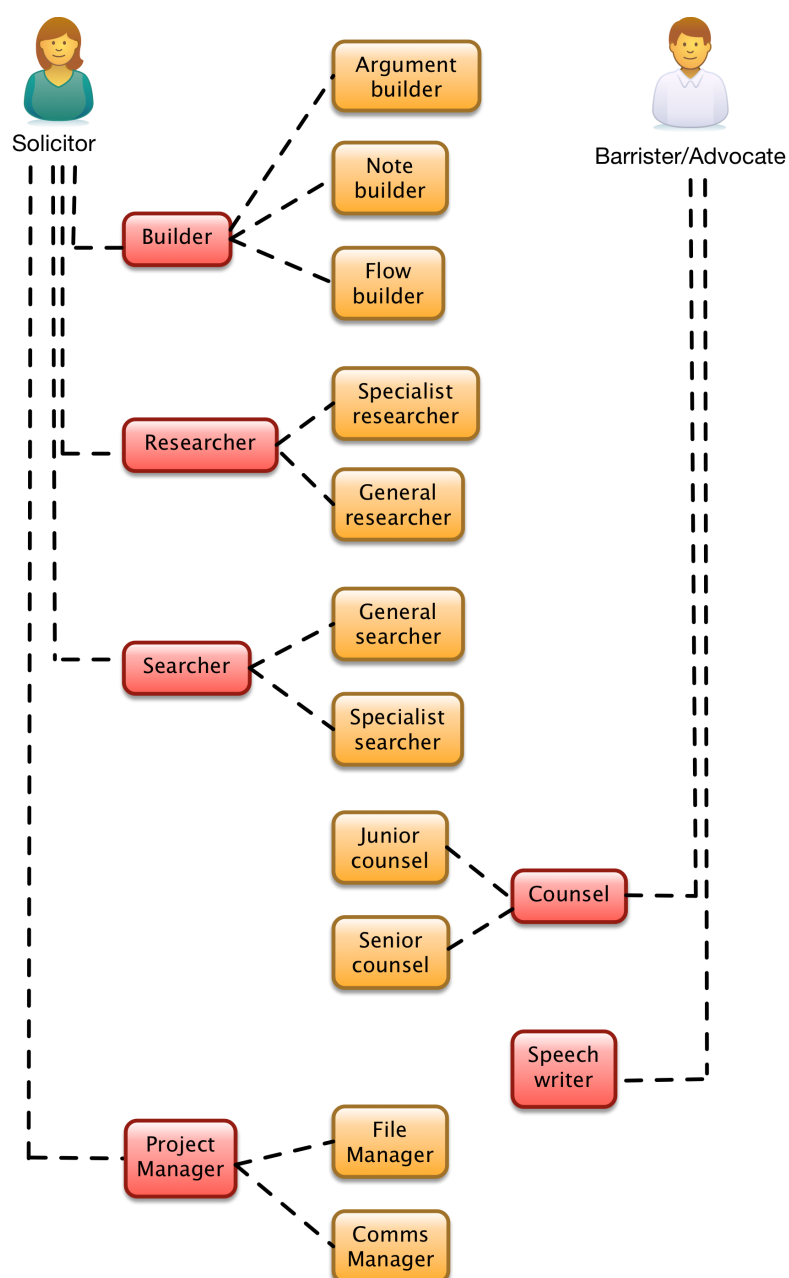


Figure A2: Work roles identified from audio analysis of the solicitor interviews

A.2 Mooting Spatial Arrangement

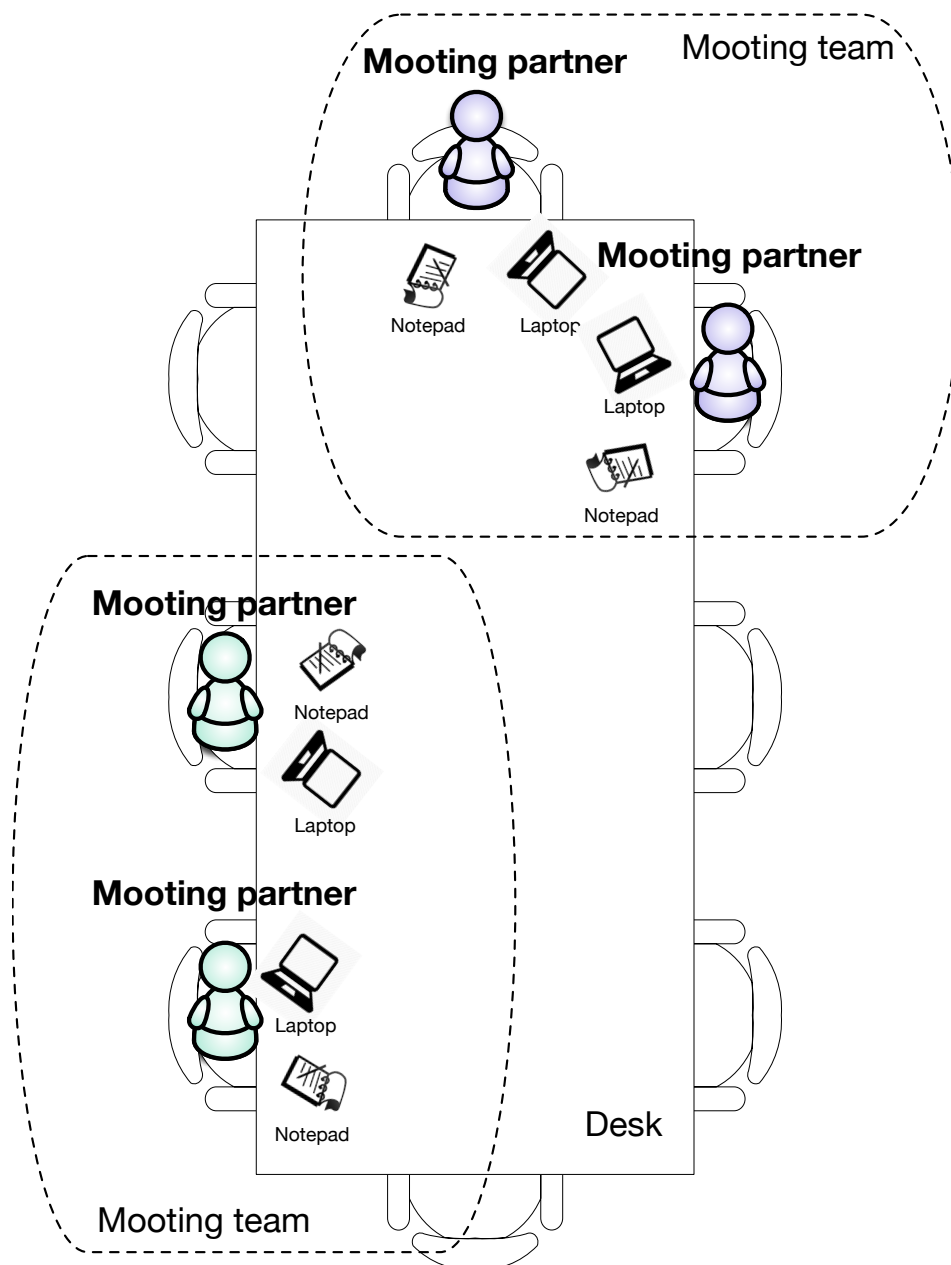


Figure A3: The physical arrangement of the mooting preparation group

A.3 Hierarchical Task Analysis

A.3.1 Reading the problem question

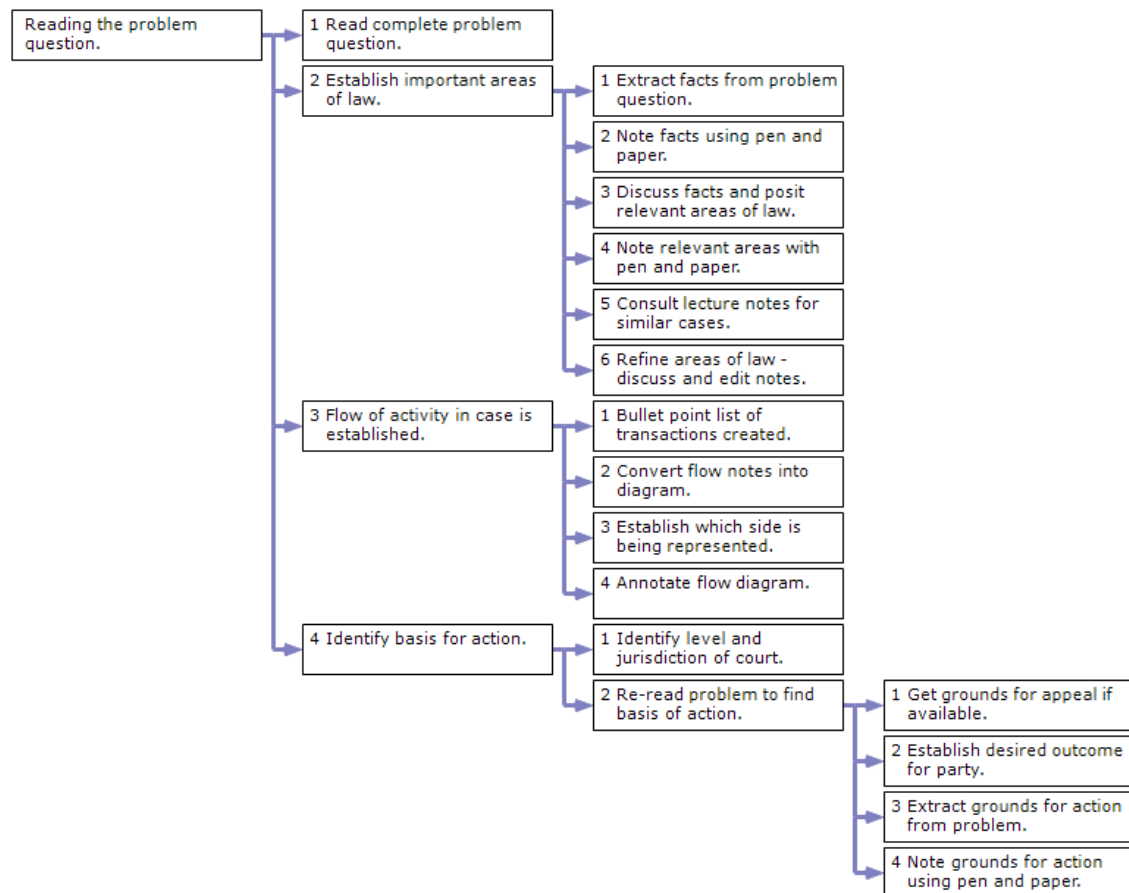


Figure A4: HTA for reading the problem question

A.3.2 Identifying seed cases

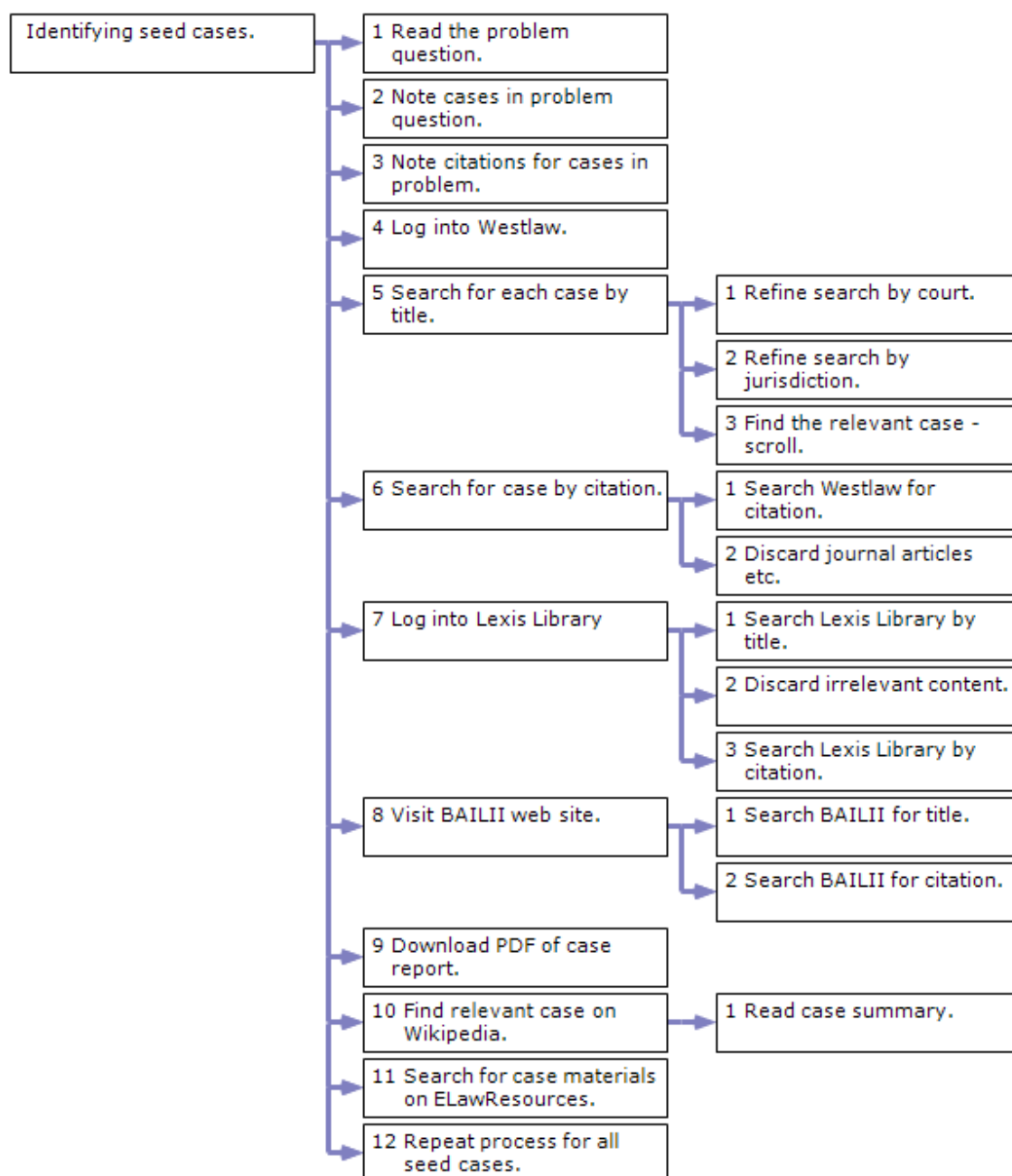


Figure A5: HTA for identifying seed cases

A.3.3 Splitting the problem question

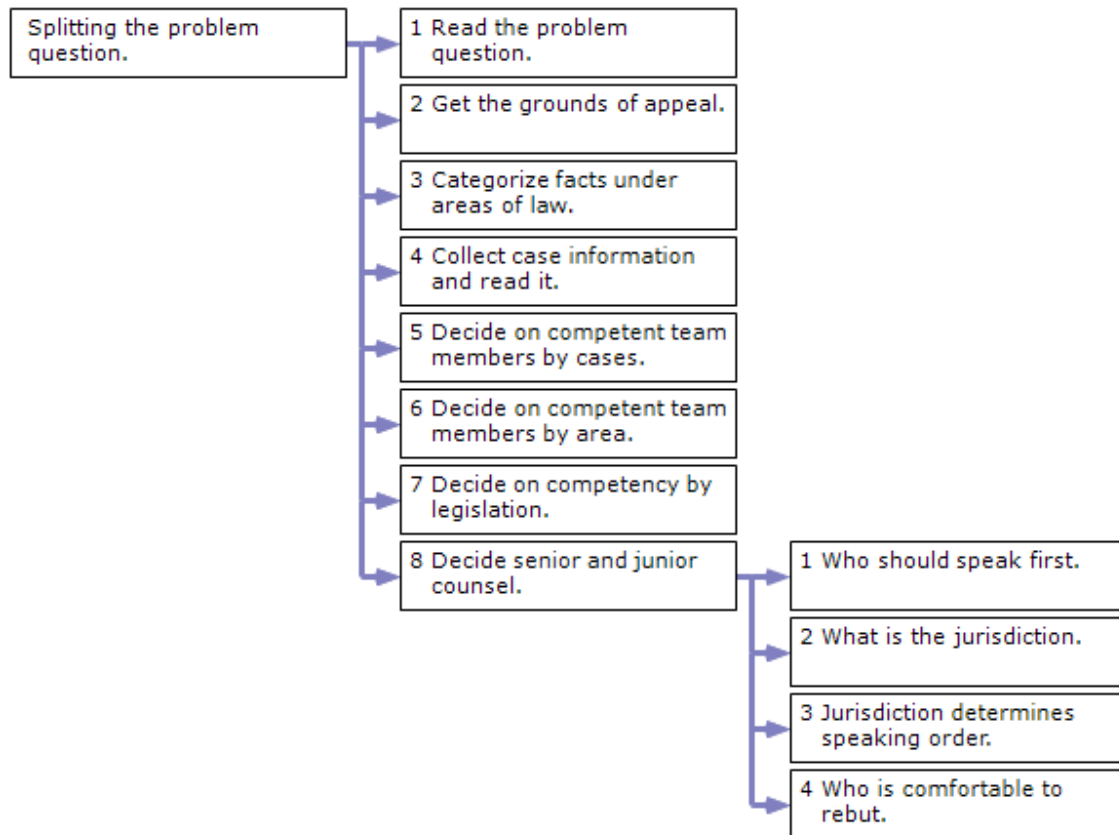


Figure A6: HTA for splitting the problem question

A.3.4 Searching for relevant cases

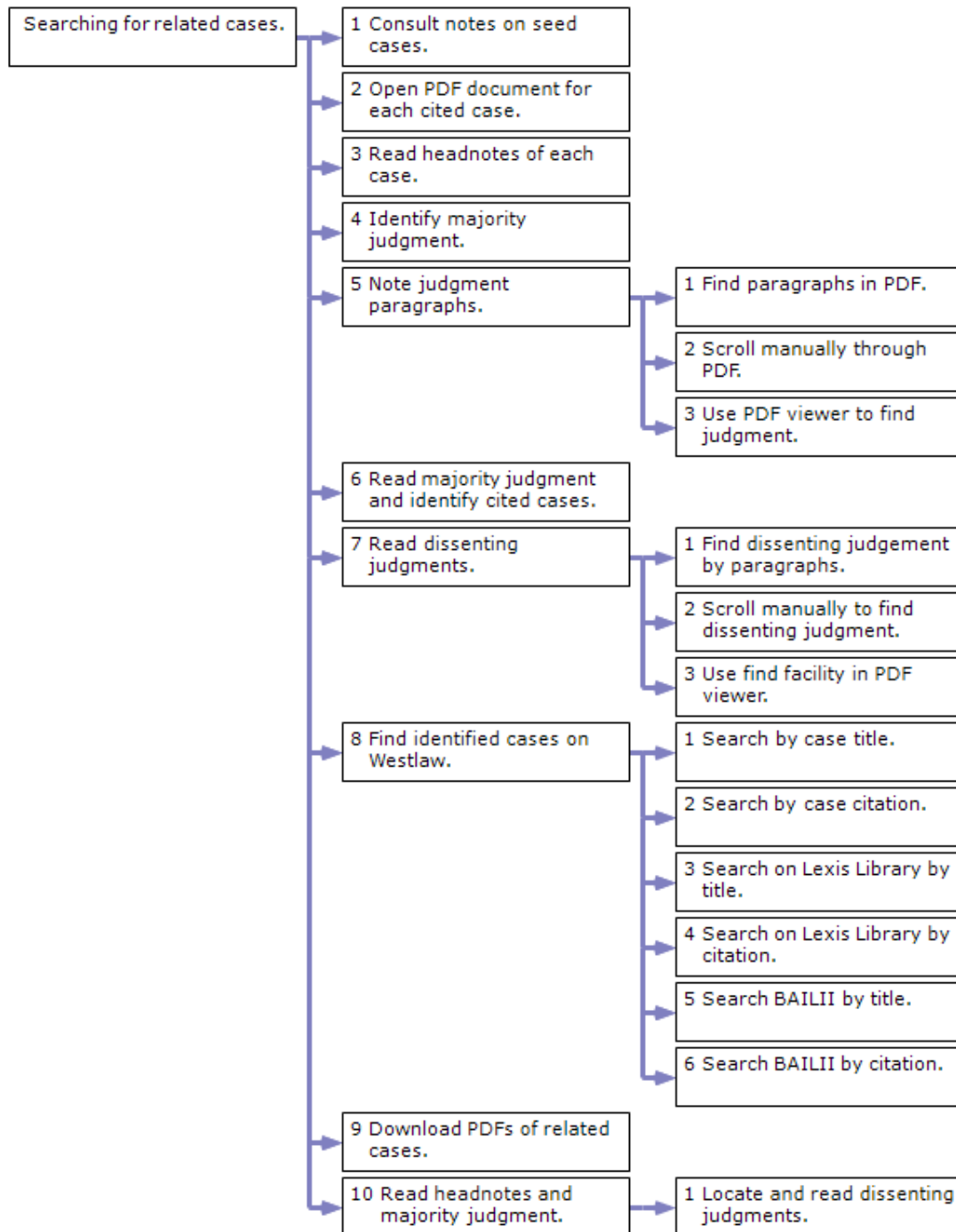


Figure A7: HTA for searching for relevant cases

A.3.5 Searching for relevant legislation

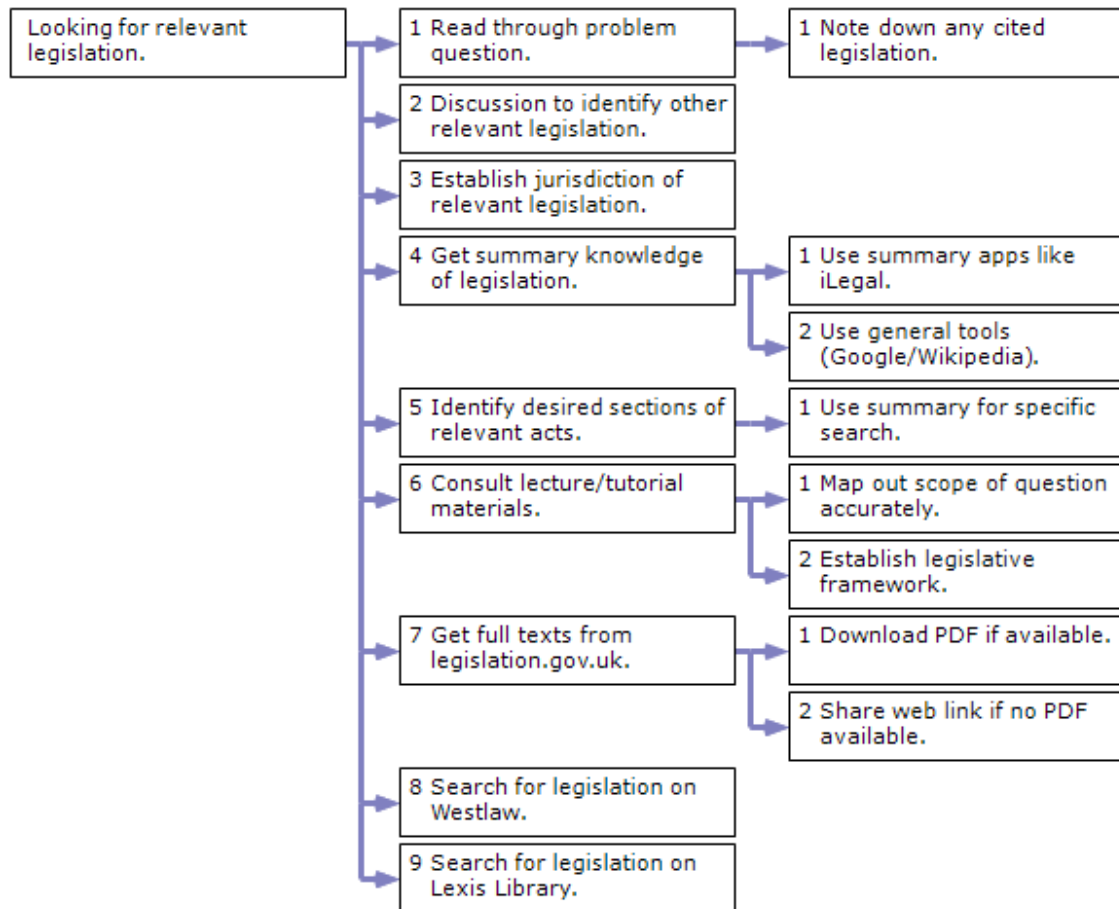


Figure A8: HTA for searching for relevant legislation

A.3.6 Identifying relevant journal articles

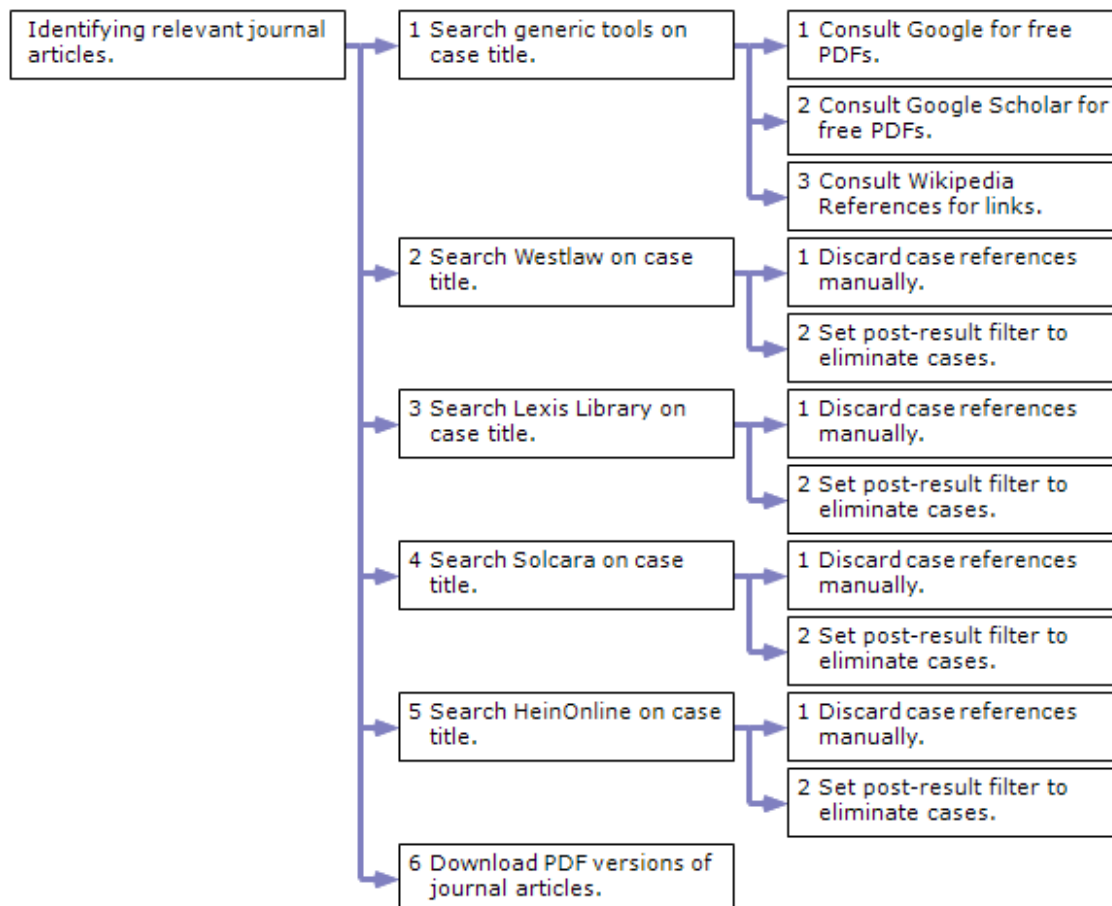


Figure A9: HTA for identifying relevant journal articles

A.3.7 Building an argument

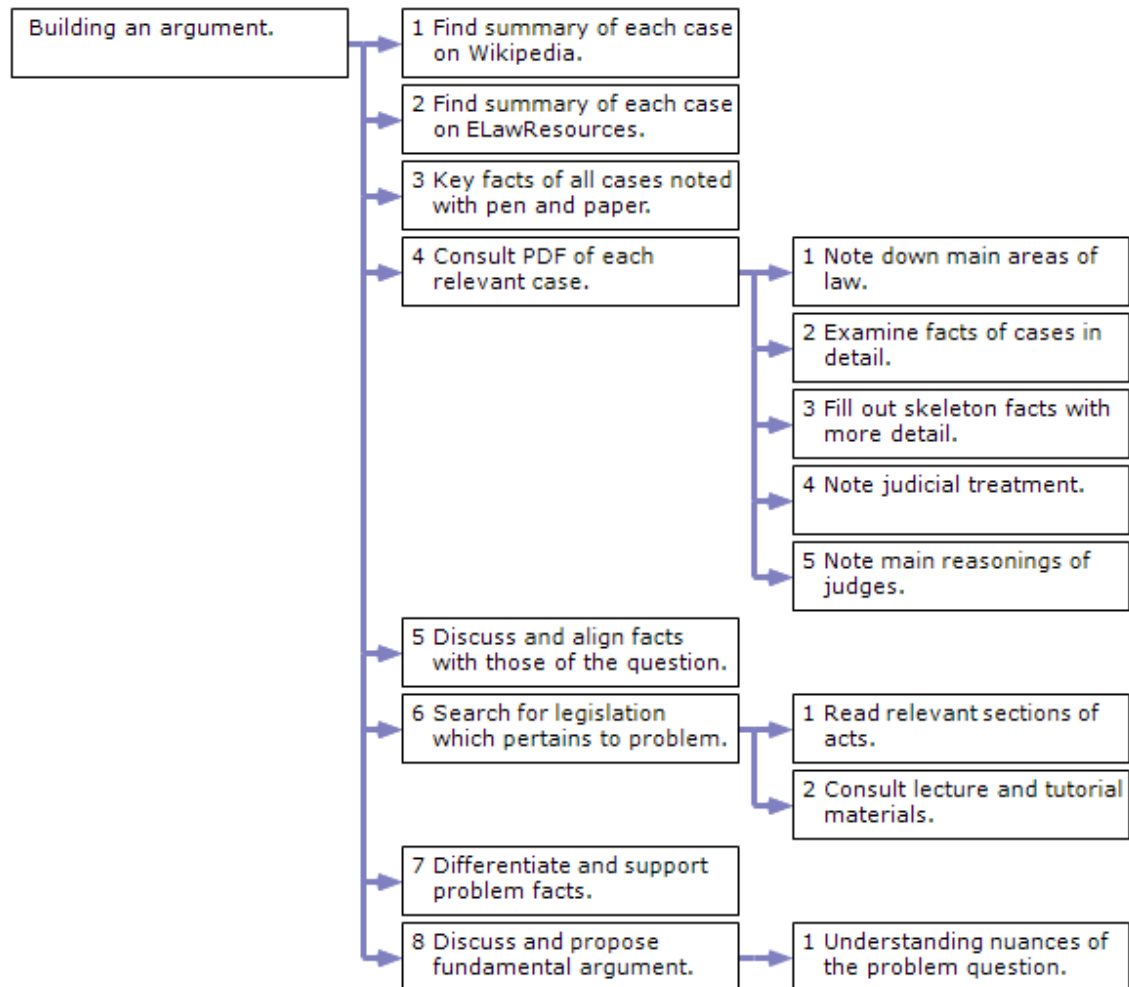


Figure A10: HTA for building an argument

A.3.8 Identifying the counter-argument

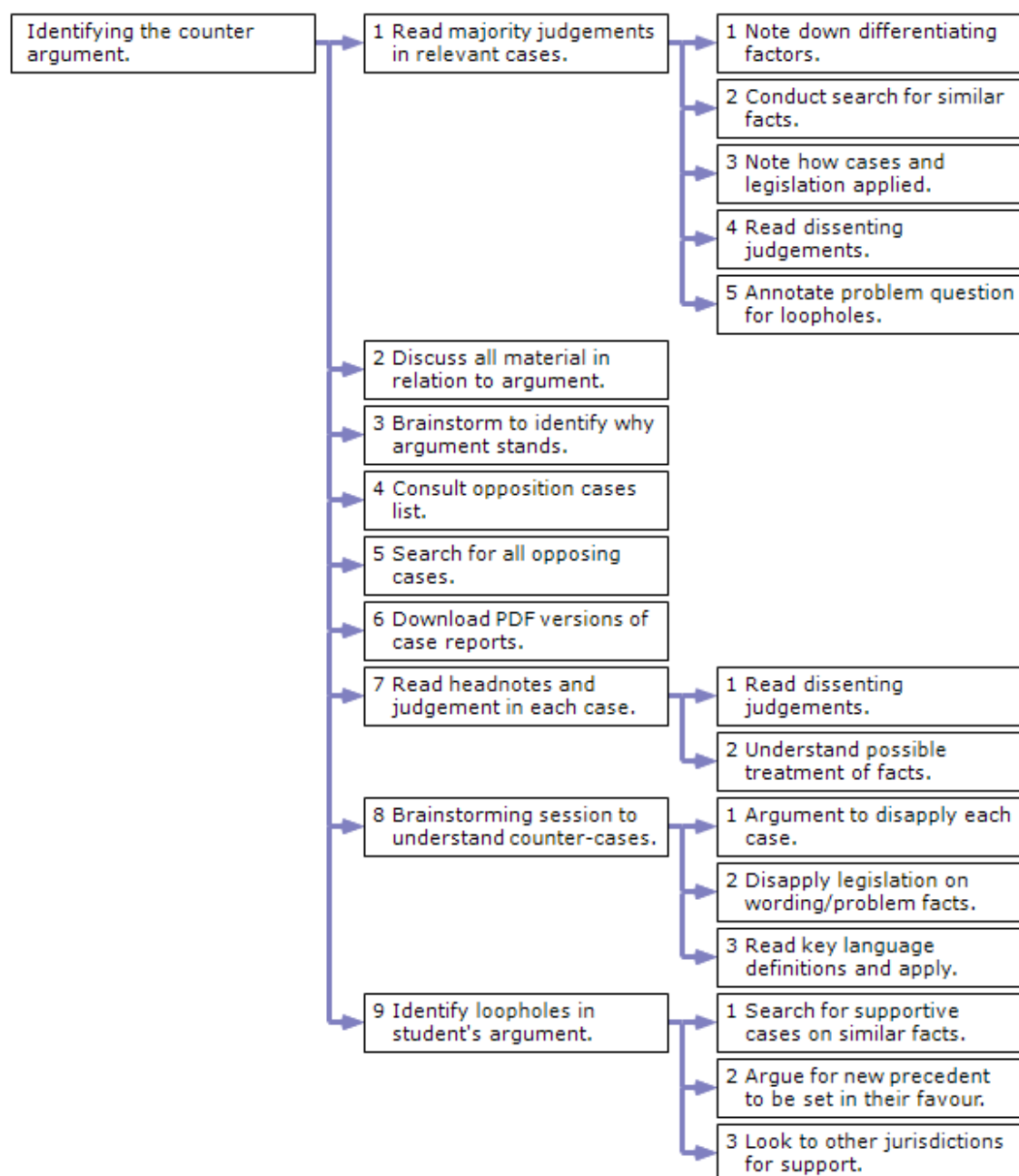


Figure A11: HTA for identifying the counter argument

A.3.9 Preparing a speech

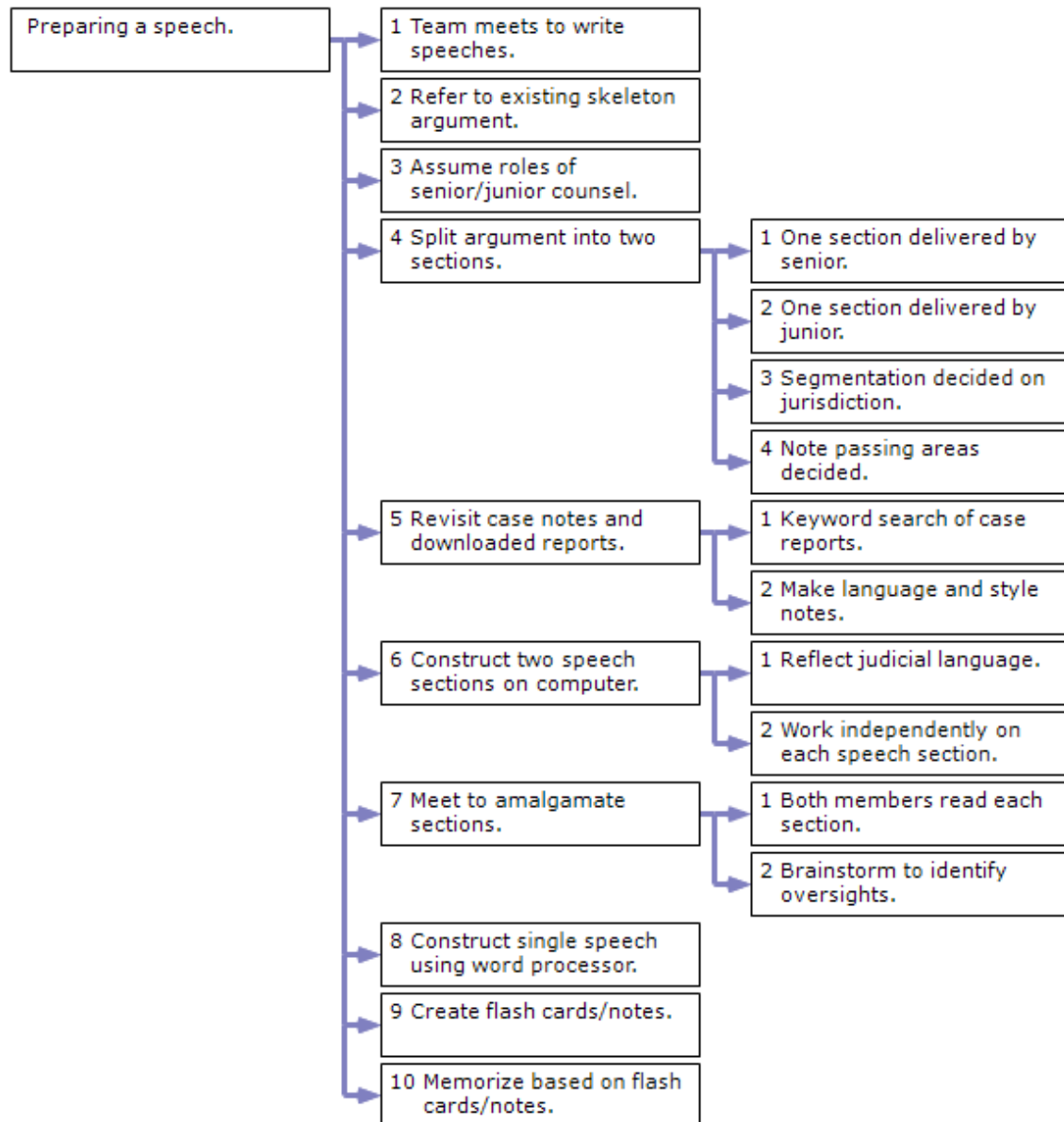


Figure A12: HTA for preparing a speech

A.3.10 Preparing a rebuttal

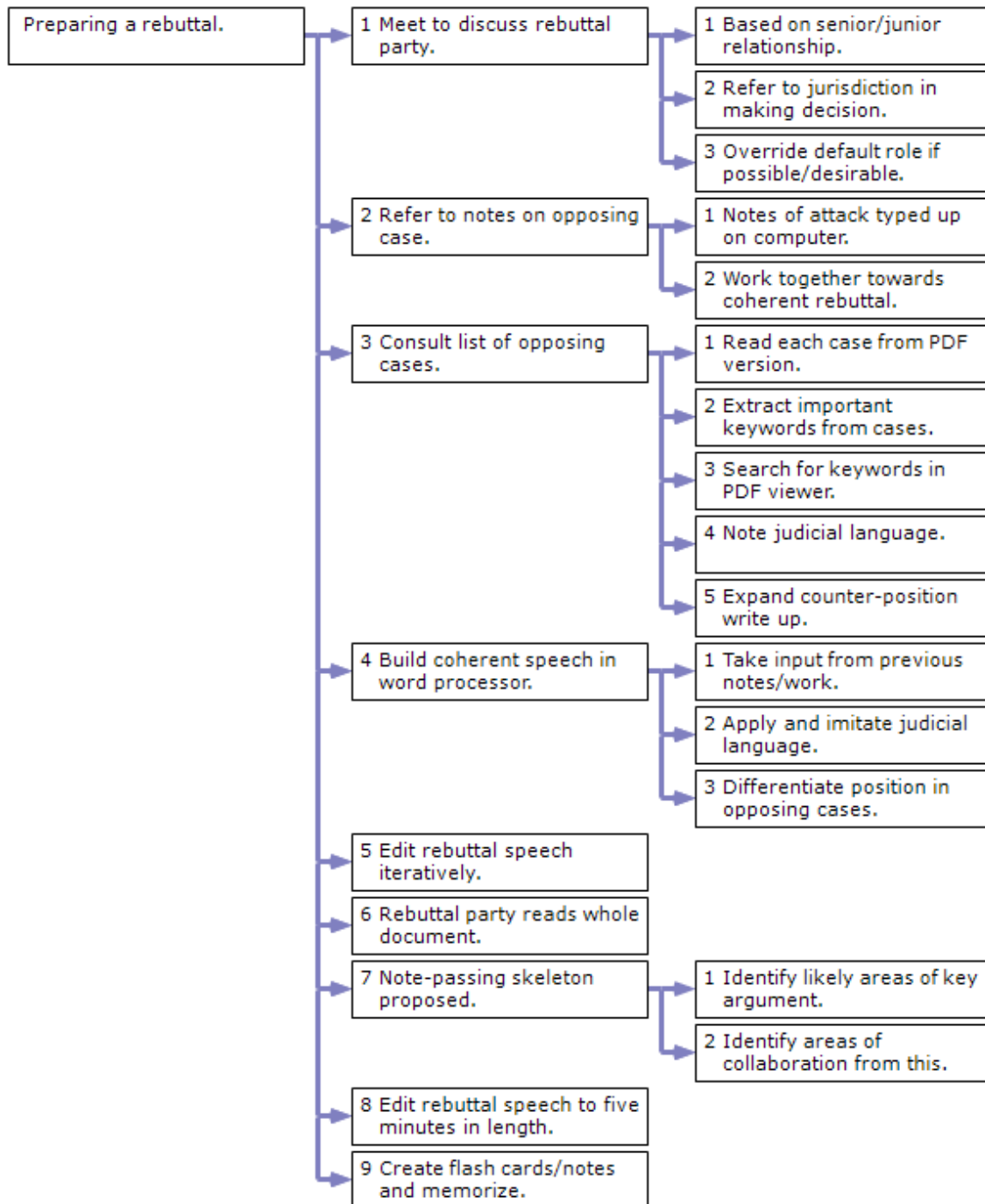


Figure A13: HTA for preparing a rebuttal

A.4 Flow diagrams

A.4.1 Reading the problem question

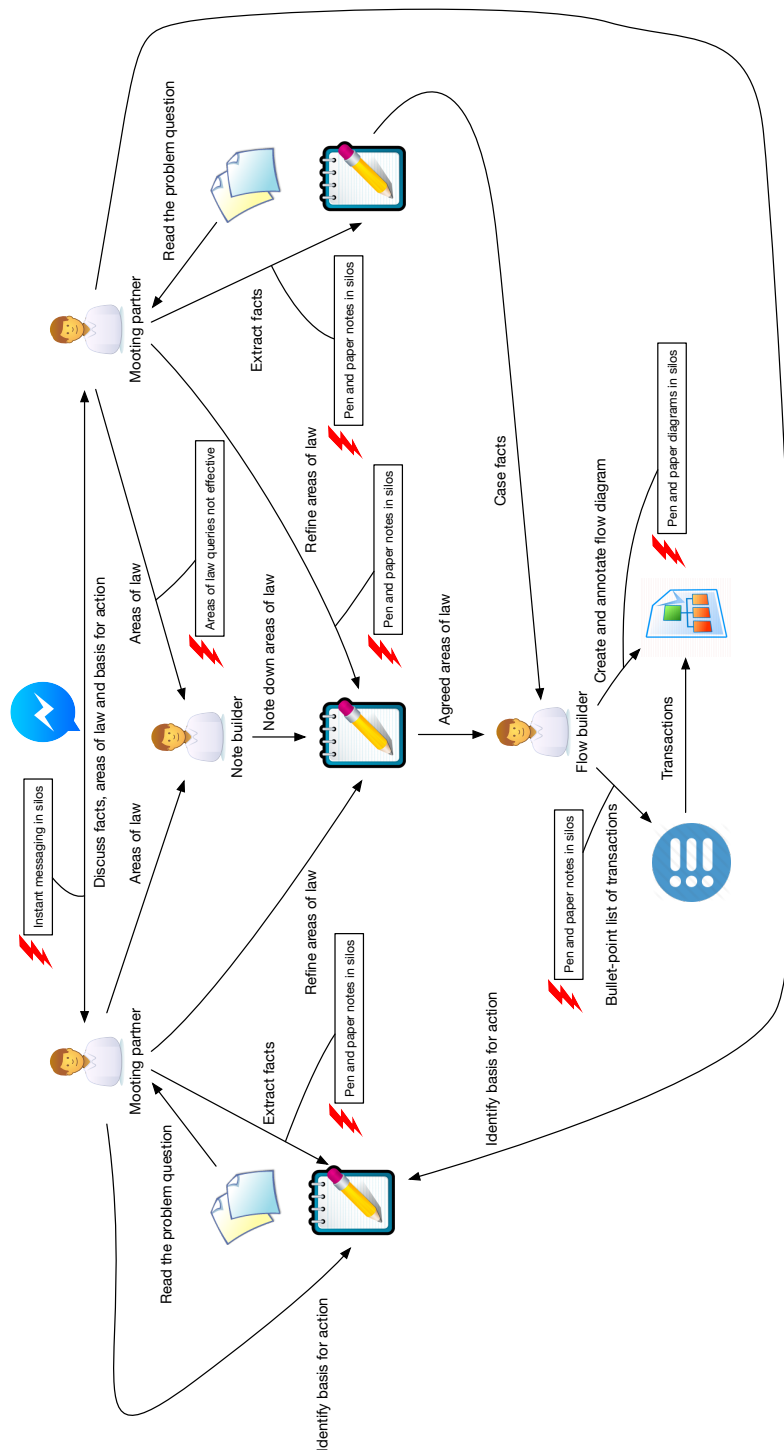


Figure A14: Flow diagram for reading the problem question

A.4.2 Identifying seed cases

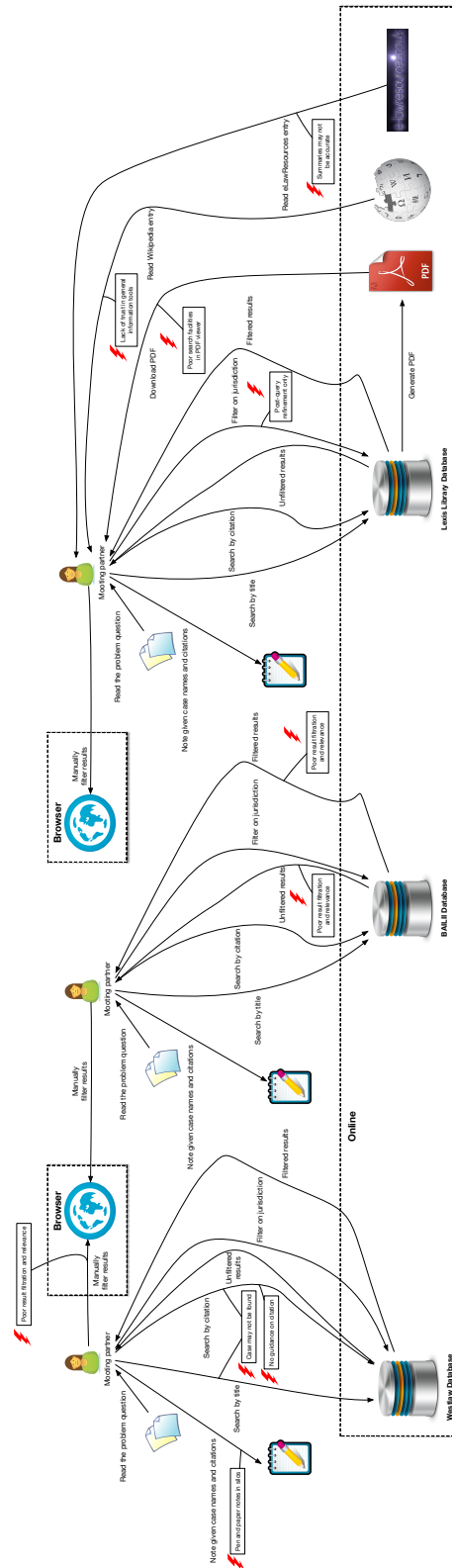


Figure A15: Flow diagram for identifying seed cases

A.4.3 Splitting the problem question

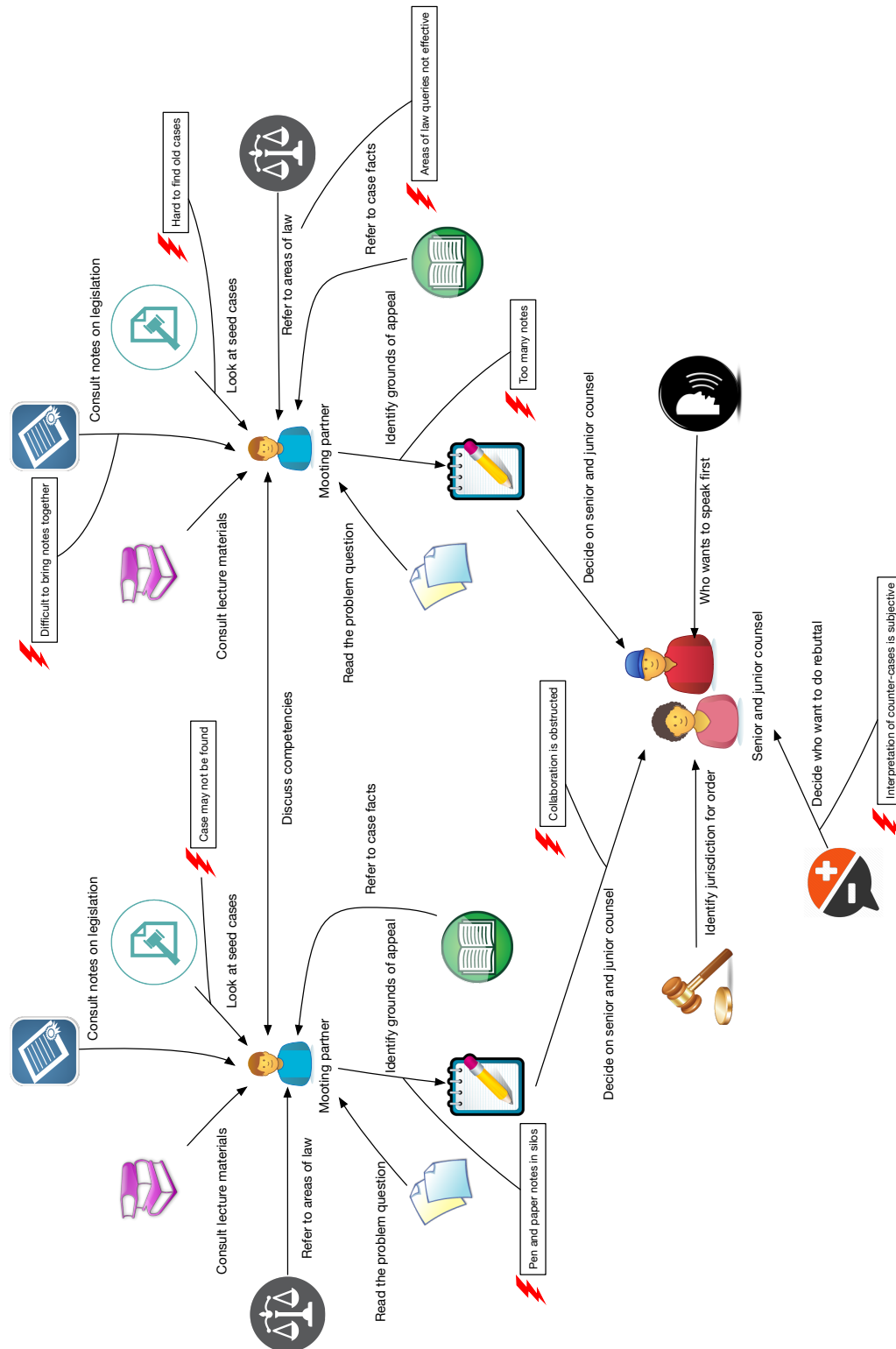


Figure A16: Flow diagram for splitting the problem question

A.4.4 Searching for relevant cases

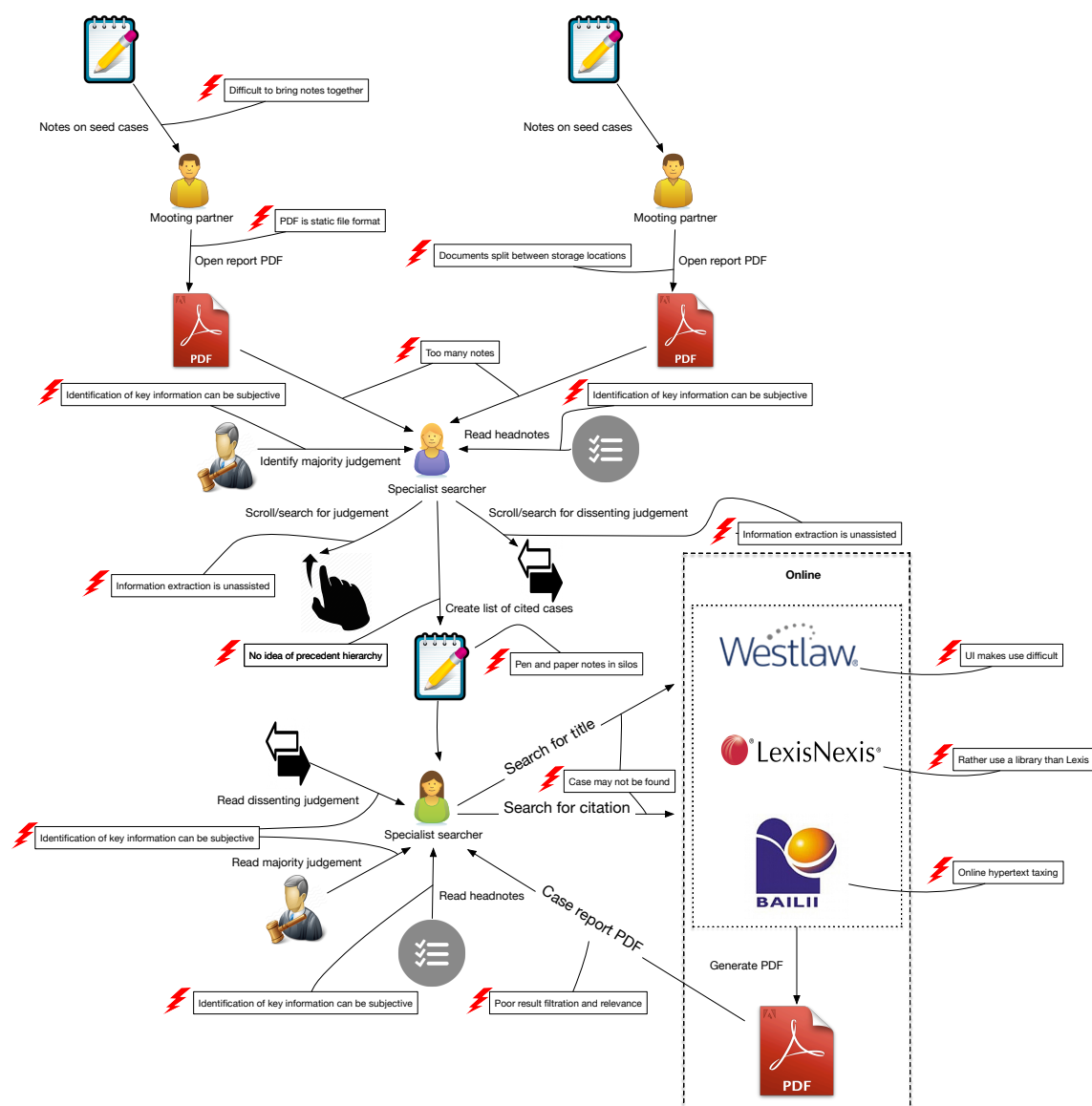


Figure A17: Flow diagram for searching for relevant cases

A.4.5 Searching for relevant legislation

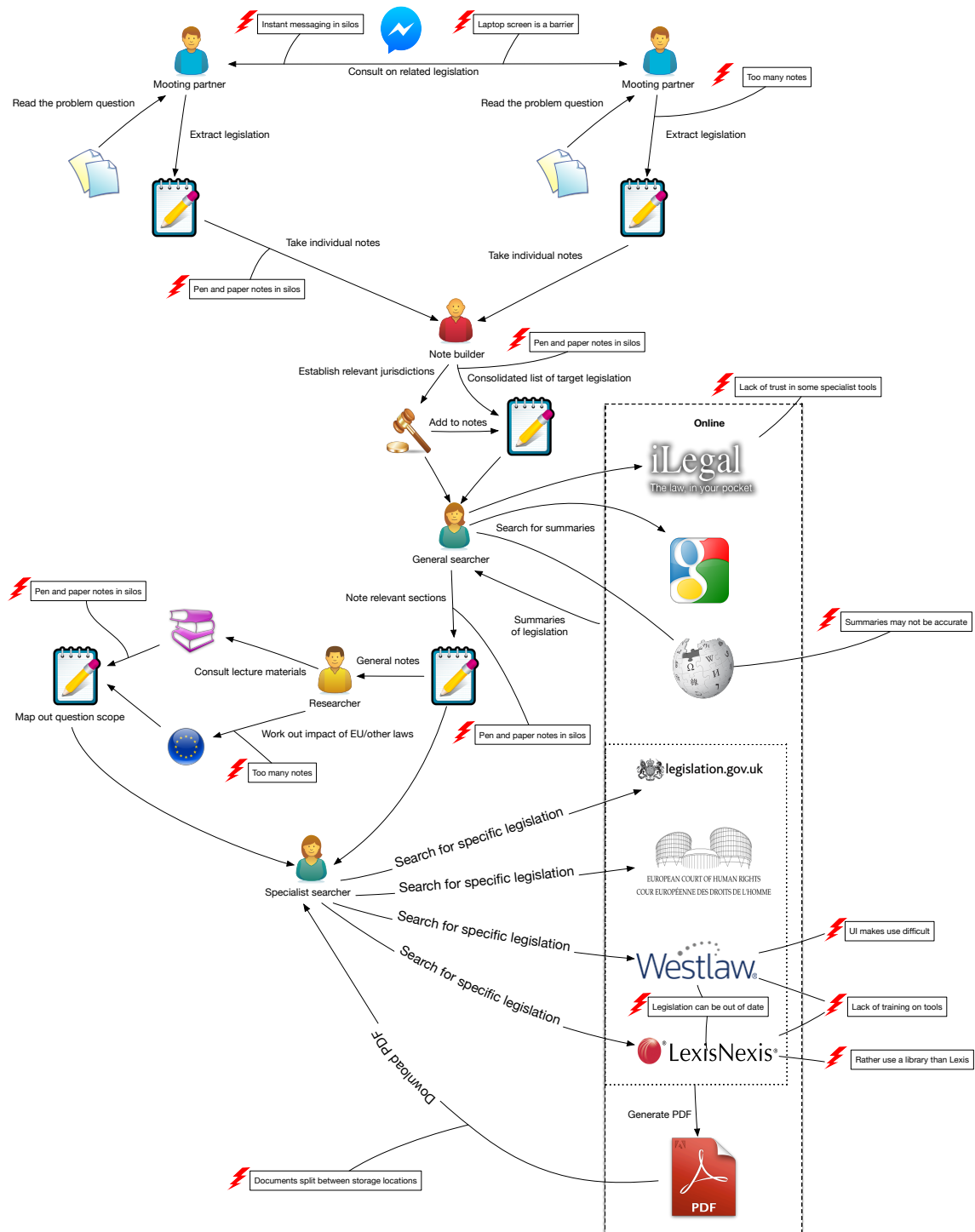


Figure A18: Flow diagram for searching for relevant legislation

A.4.6 Identifying relevant journal articles

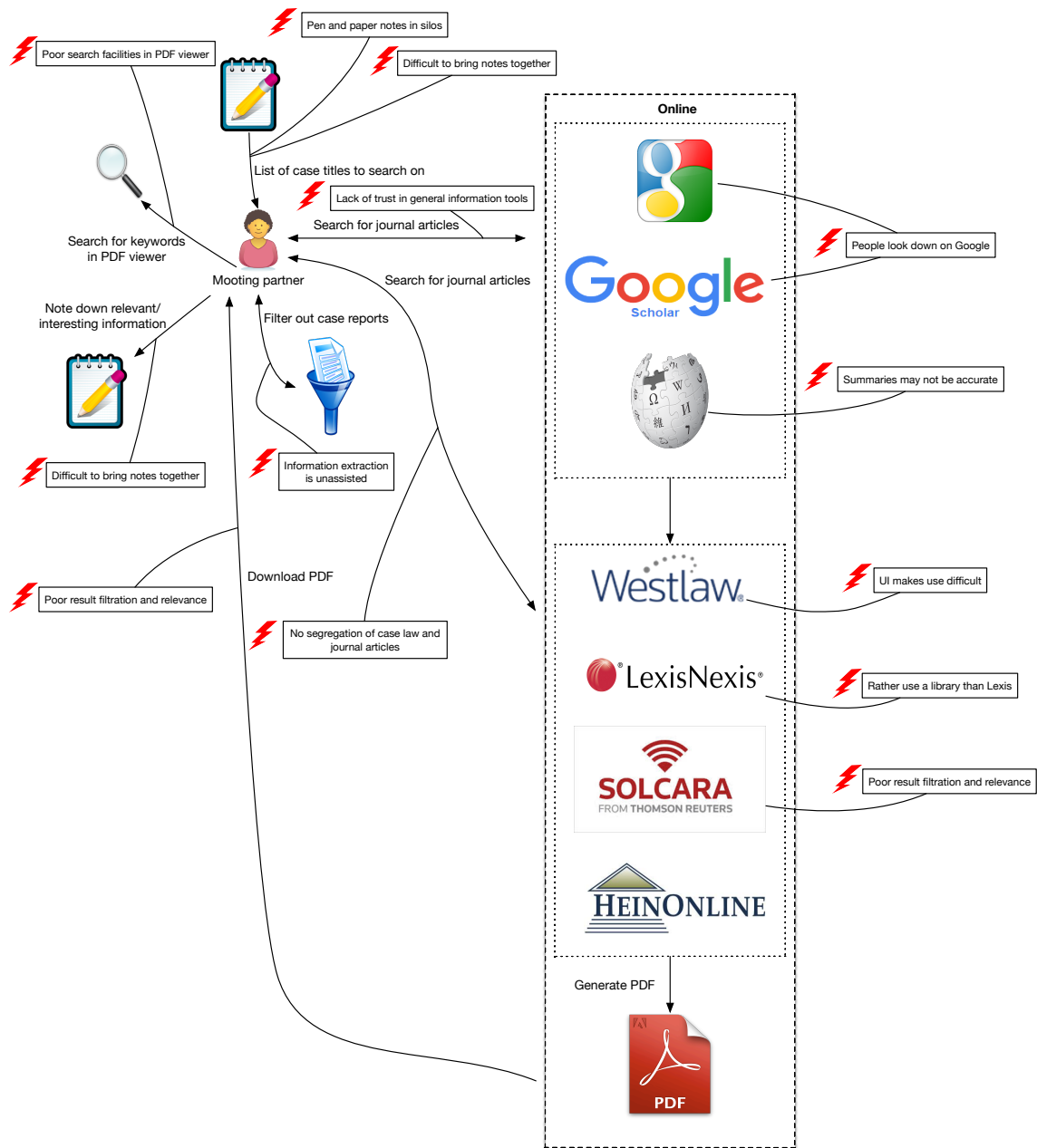


Figure A19: Flow diagram for identifying relevant journal articles

A.4.7 Building an argument

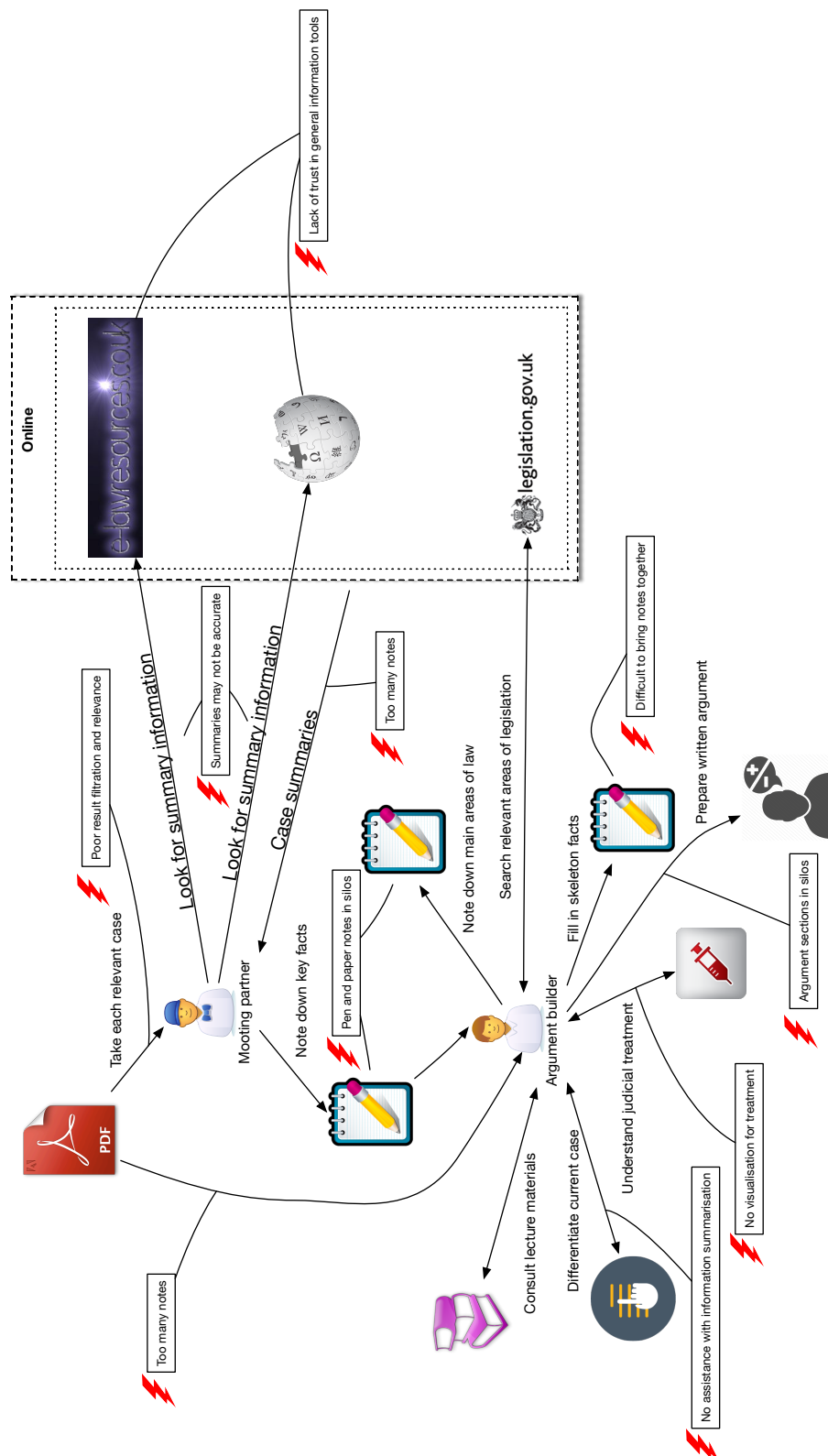


Figure A20: Flow diagram for building an argument

A.4.9 Preparing a speech

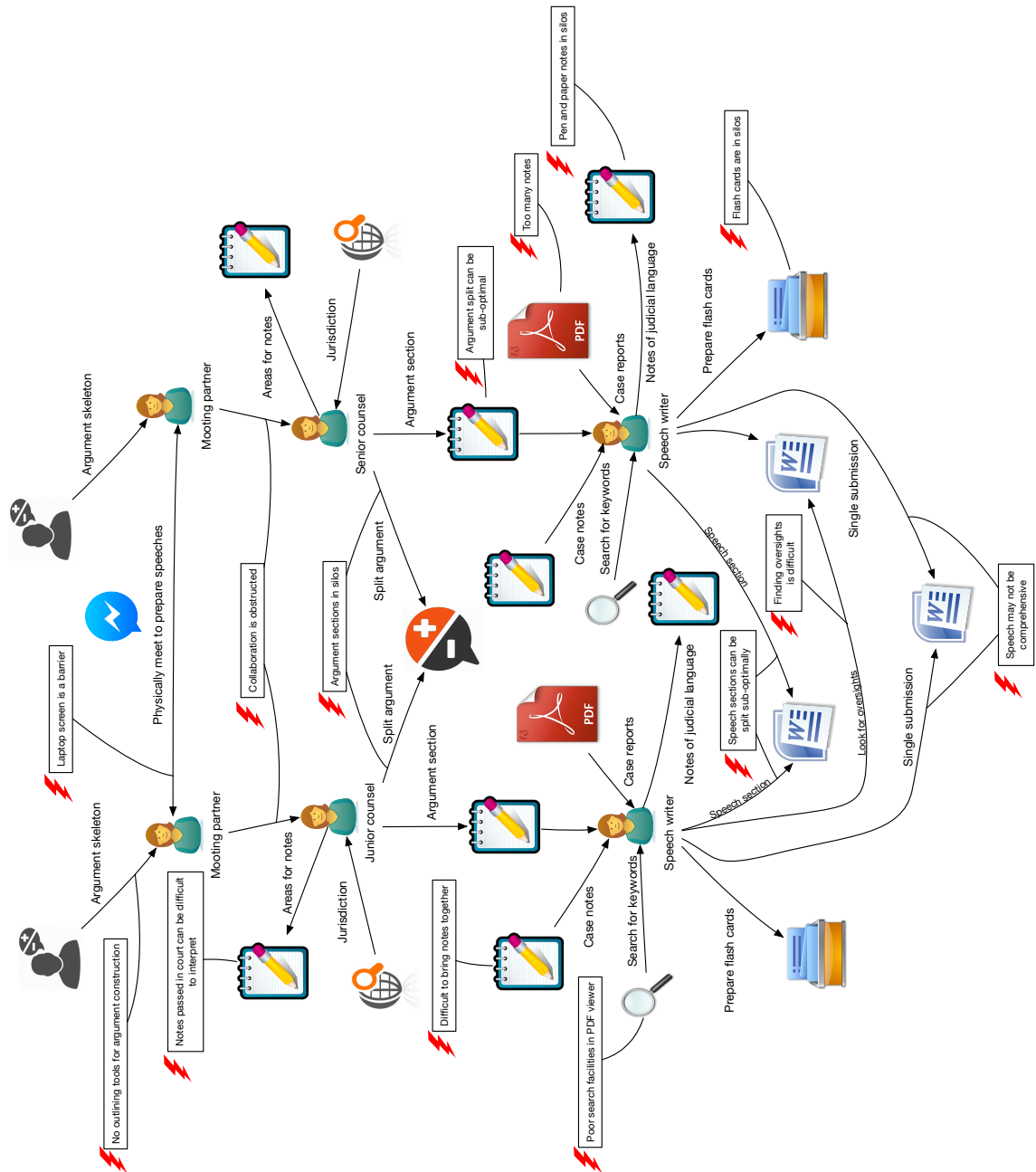


Figure A22: Flow diagram for preparing a speech

A.4.10 Preparing a rebuttal

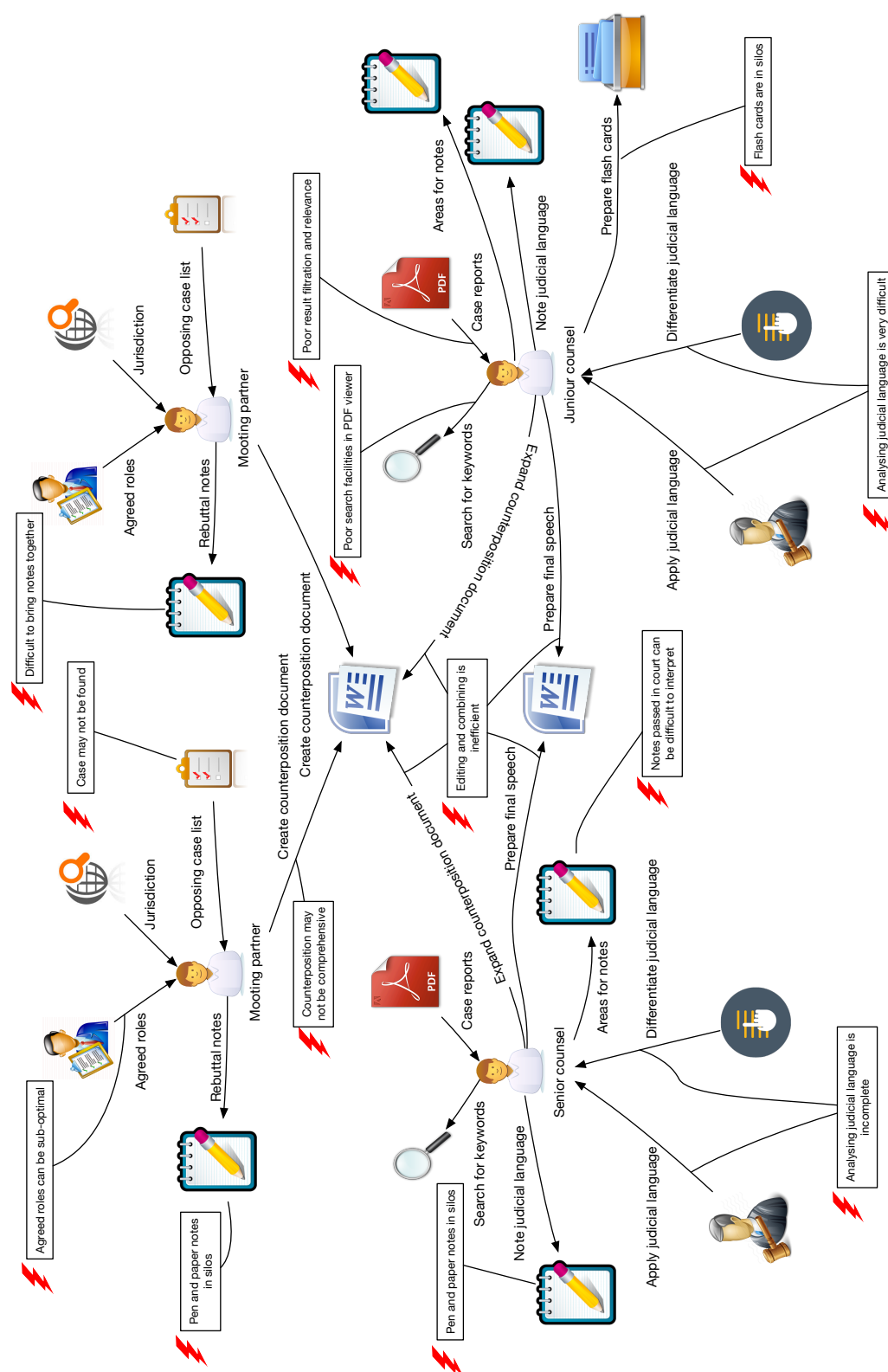


Figure A23: Flow diagram for preparing a rebuttal

A.5 Artefact model

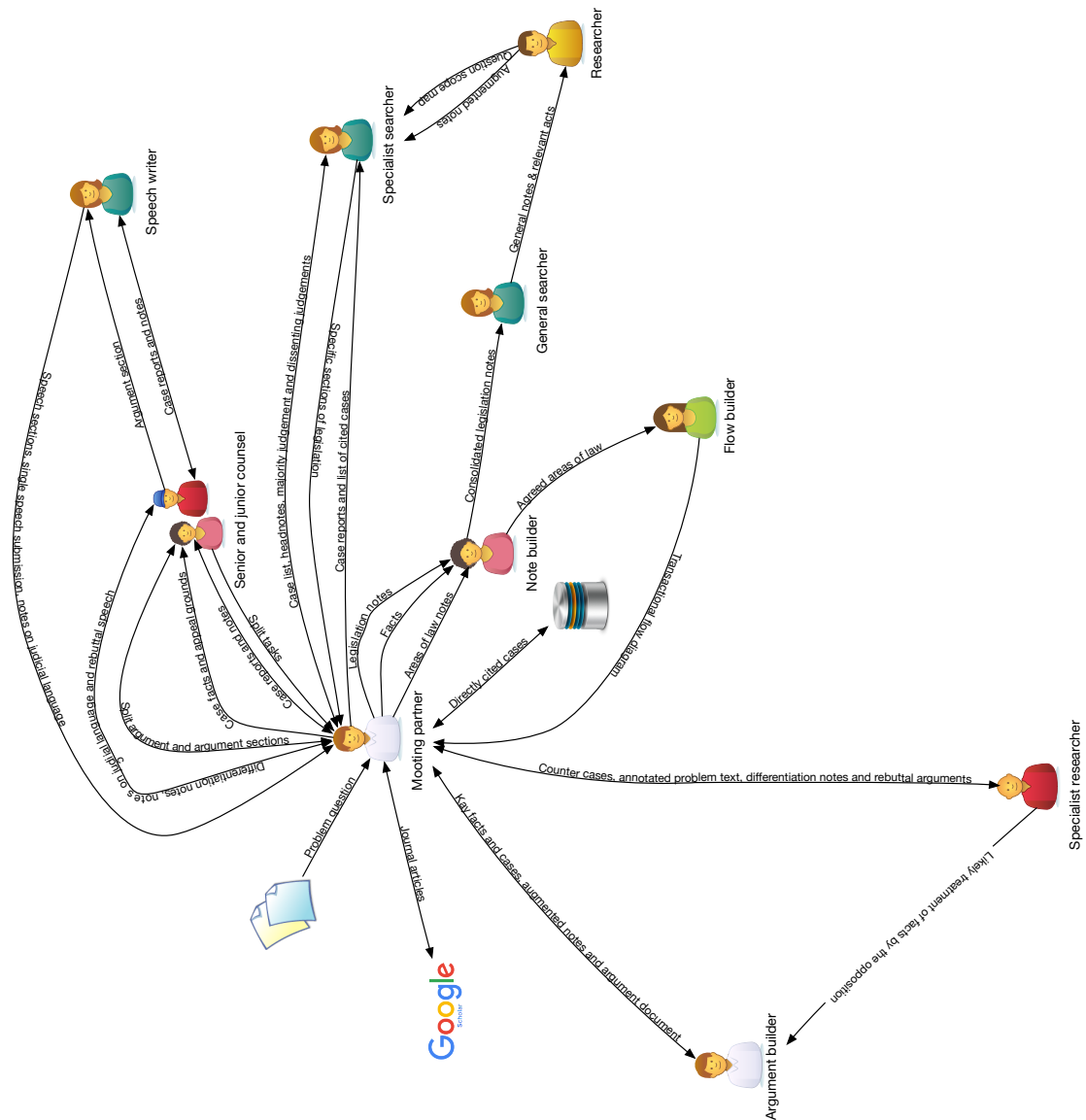


Figure A24: Artefact diagram between work roles

A.6 Social model

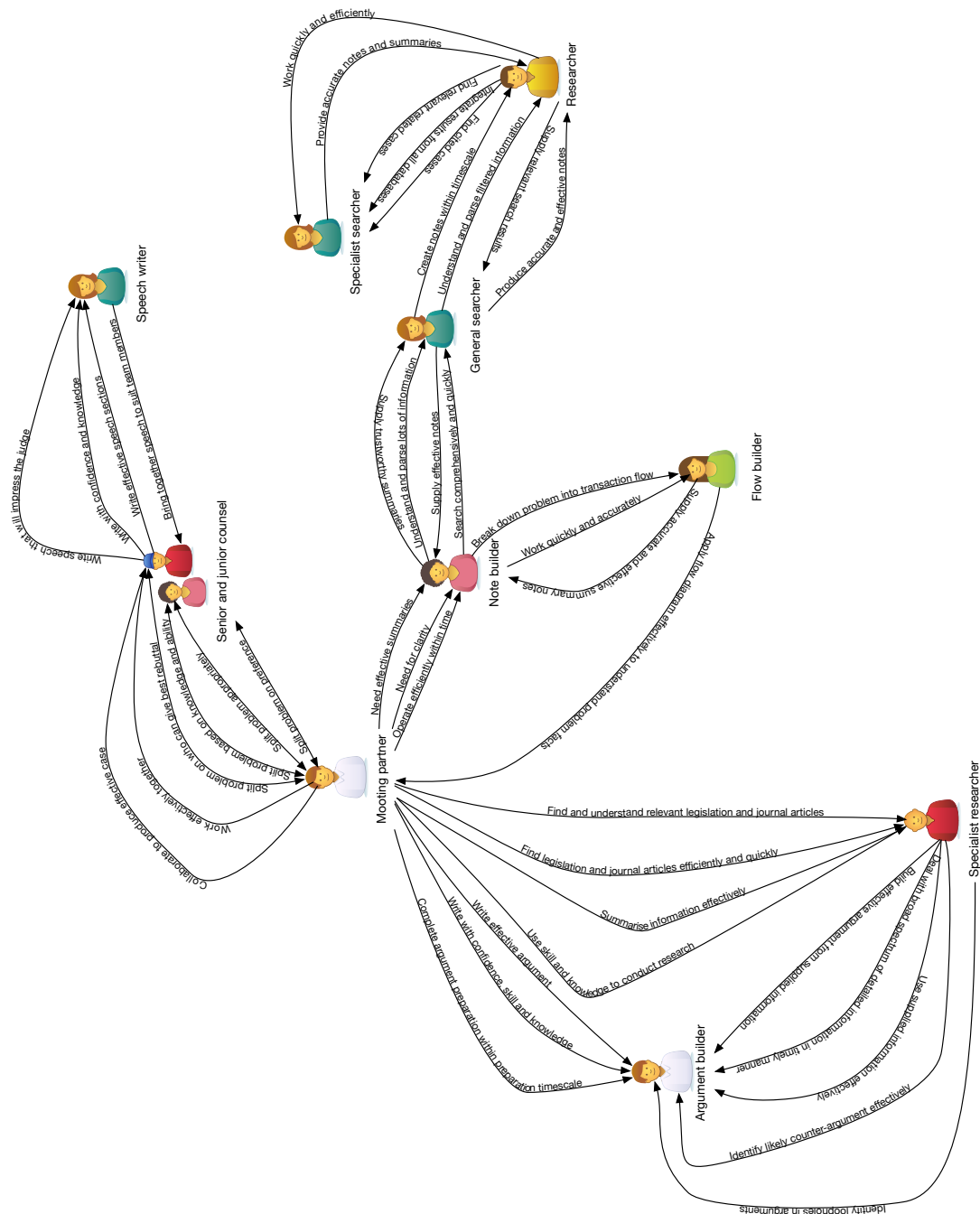


Figure A25: Social diagram showing interactions between mooring actors

A.7 Affinity diagrams

A.7.1 Affinity diagram A

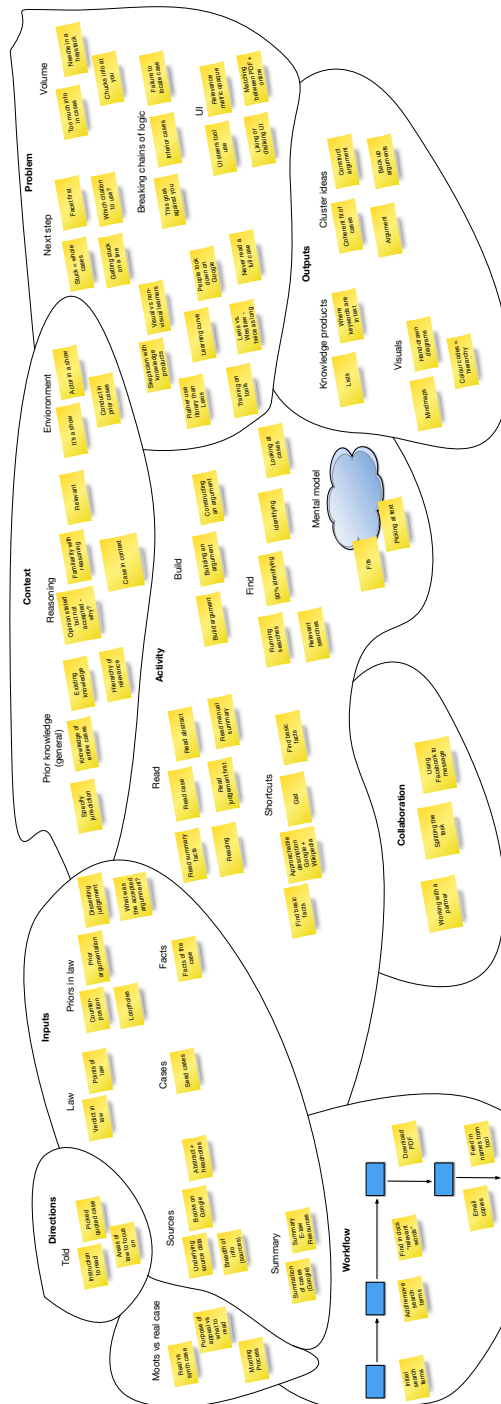


Figure A26: Affinity diagram A

A.7.2 Affinity diagram B

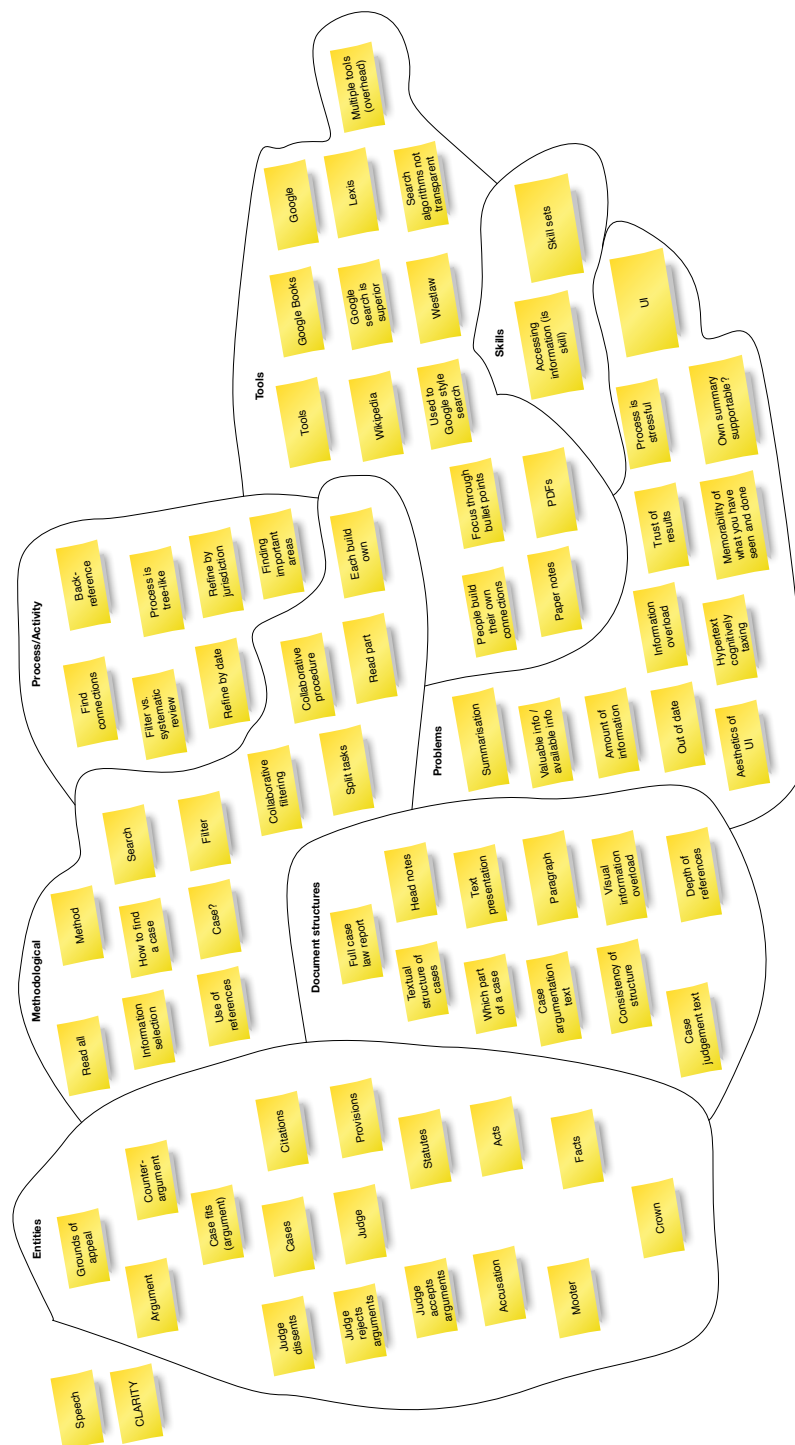


Figure A27: Affinity diagram B

A. SETTING THE SCENE: CONTEXTUAL INQUIRY, INTERVIEWS AND SURVEY

A.7.4 Consolidated affinity diagram

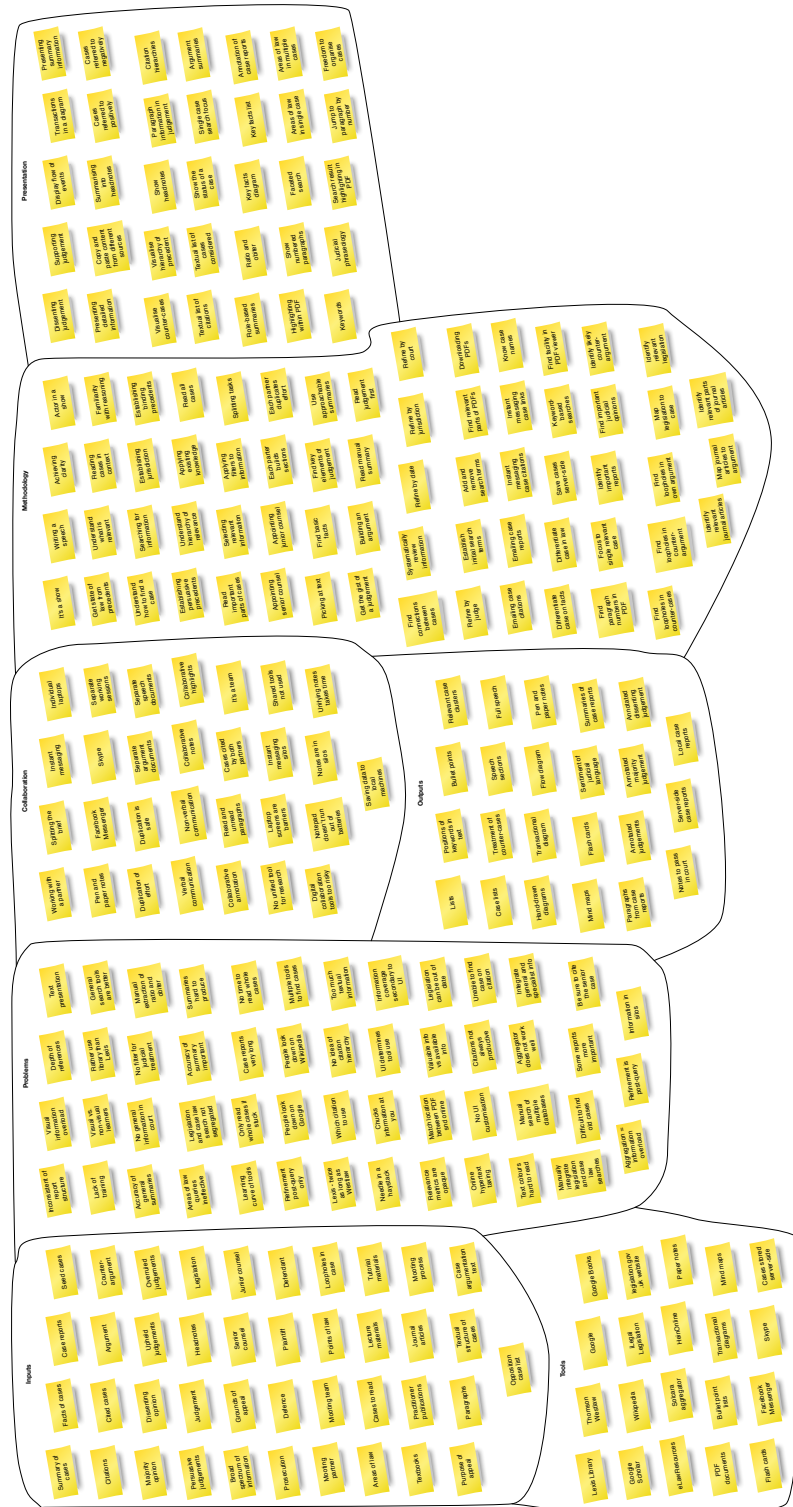


Figure A29: Consolidated affinity diagram from the three initial iterations

A.8 Barrier severity matrix



Figure A30: Severity matrix for identified barriers to effective working practice

A.9 Mooting interview questions

General introductory questions

- How much time do you spend identifying relevant legal precedents during the process of preparing for a moot?
- To what extent do you use precedents that you have memorised from teaching and tutorials as opposed to precedents that you must first discover?
- What is your main purpose in using computer-based legal information tools when preparing for a moot?
- Which are your favourite and least favourite information discovery tools?
- Do you use an aggregator in order to bring together search results from multiple legal discovery platforms?

Establishing the importance of the user interface

- To what extent does the user interface of different tools inform your preference for one platform as opposed to another?
- What is the main factor that makes you like or dislike a given tool?
- How much time on average will you spend using a computer to conduct research for a moot?
- How much of the time do you spend learning to use the tool as opposed to finding relevant precedents or other information?

Access to the full text of cases

- To what extent does access to the full text of case law reports inform your research?
- Do you often read or need access to the full text of case law reports or do you base your analysis of cases predominantly on head notes?
- How well does your favourite information discovery tool provide access to the full text of case law reports?

- Are the keywords highlighted in search results relevant and accurate enough to give you a quick understanding of the main judgement in a case or do you typically have to explore the text of the report further?
- Do you find textual search results from your favourite information discovery tool easy to digest?

Visualisation of search results

- Could the information provided by your preferred case law discovery tool be presented differently to promote discovery and understanding?
- Are the search results that you have obtained relevant, comprehensive and timely?
- What improvements would you like to see in the way in which the discovery platform presents search results?
- Do you feel that the search pathway available to you in your favourite information discovery tool is clear?
- Do you feel that it is easy to refine searches that are conducted using the tool so that you are not left at a dead end during the research process?
- What additional information could the platform usefully give you that it currently does not?

Working in groups

- How easy is it to split work tasks between groups of people using the computer-based tools that are available to you?
- Do you typically identify one person in the group who will be responsible for researching relevant precedents?
- How do you typically divide responsibilities for moot preparation between the members of the group who are involved?

The relationship between citations and case text

A. SETTING THE SCENE: CONTEXTUAL INQUIRY, INTERVIEWS AND SURVEY

- How easy is it to view citations for cases in your favourite information discovery tool?
- What is the process of moving from a view of case citations to a view of the relevant search result hits in the full text of the case law report that you are interested in?
- To what extent does the information discovery tool allow you to view search hits which occur within a particular case law report as opposed to individual search hits which occur across the returned collection as a whole?
- What are the critical search facets that you are most interested in when you search for precedents (i.e. the Court the case was heard in, the year the case was heard in, the judge who delivered the main judgement, related cases which build upon or invalidate a particular judgement and so on?)?

The relevance and importance of mooting

- What are the benefits to you of participating in moots and mooting competitions?
- What first motivated you to get involved in mooting and mooting competitions?
- To what extent do you think the process of preparing for a moot is similar to the process of preparing for a real legal case and argument?
- What are the main ways in which mooting improves your understanding of the law and the legal process?

Specific questions for the staff cohort

- How have electronic legal information tools improved teaching and facilitated learning?
- What are the disadvantages to using and promoting the use of electronic tools compared to traditional printed digests of cases?
- To what extent does the use of computer-based platforms enhance the ability of a student to understand and explore the implications of legal cases?

- From a teaching perspective, how would you like to see case law discovery platforms improved?
- Do you think that the visualisation of search results is an area where change is a priority at the moment?
- To what extent do legal search tools provide better access to relevant information than standard search platforms like Google or Bing?
- Are the costs of licensing legal tools for use in the University justified?
- To what extent do you recommend that your students make use of free legal tools and databases such as the British and Irish Legal Information Institute's online search facility?
- What are the advantages of free tools over paid-for alternatives?
- In what ways are free tools less effective than the various commercial offerings that are available?
- What are the biggest improvements that could be made to legal information discovery tools to increase their effectiveness and relevance for learning and teaching?

A.10 Solicitor interview questions

General introductory questions

- How much time do you spend identifying relevant legal precedents during the process of preparing for litigation?
- In preparing for litigation, what main work functions do computer-based tools support?
- In preparing for litigation, what work functions could be better supported by computer-based tools and why?
- If you could design a new computer-based tool to help you in your work, what would it do and what key features would it have?
- How much time do you spend using general purpose computer-based tools in your work - i.e. Google search, Microsoft Word, email etc.?
- How much time do you spend using specialist tools for lawyers - i.e. Lexis Library, Westlaw, JustCite, Solcara?
- Do you use tools which integrate legal information discovery and drafting tools into Microsoft Word or other general software - i.e. Lexis for Word?

Working in groups

- What are the main collaborative activities in preparing for litigation?
- How much time do you spend working in groups when preparing for litigation?
- Is collaboration typically synchronous or asynchronous?
- How is the preparation process for litigation typically split between different people?
- Who are the stakeholders in the preparation process - associates, paralegals, lawyers, barristers etc.?
- Is legal information discovery typically delegated to associates and paralegals?

- How are arguments and cases developed from the work outputs of multiple people?
- Do you use particular tools to enable collaboration with computers and digital data - Dropbox, Skype, SharePoint etc.?
- Are there specialist legal tools which you use in order to facilitate collaboration?
- How do you deal with duplication of effort in the case preparation process?
- Is duplication of effort a good or a bad factor in case preparation?

Establishing the importance of computer-based tools

- To what extent does the user interface of different computer-based tools inform your preference for one platform as opposed to another?
- What is the split of time in the preparation process between using a computer and performing offline tasks?
- Do you use a computer to take notes as you work?
- Do you use pen and paper to take notes?
- If you use pen and paper to make notes, why do you do this instead of using a word processor or tools like Microsoft OneNote?
- What is the lifespan of paper-based notes that you take as you work?
- If you could design a computer-based tool to help with preparations for litigation, what would its 'killer feature' be?

Access to the full text of cases

- To what extent does access to the full text of case law reports inform your preparations?
- Do you often read or need access to the full text of case law reports or do you base your analysis of cases predominantly on head notes, judgements and case summaries?

A. SETTING THE SCENE: CONTEXTUAL INQUIRY, INTERVIEWS AND SURVEY

- How well do computer-based legal information tools facilitate access to the full text of case law reports?
- Are the keywords highlighted in search results relevant and accurate or do you typically have to explore the text of the report further?
- Do you find textual search results from information discovery tools easy to digest?
- To what extent do you identify with a problem of information overload when using computer-based legal information tools?
- How important is the language used by judges in identifying relevant precedents or forming arguments in novel cases?
- Would tools which allow linguistic analysis of case reports be useful?

Information discovery

- Could the information provided by case law discovery tools be presented differently to promote effectiveness and understanding?
- To what extent do you feel that legal information search is text-heavy?
- Do you feel that it is easy to refine information searches that are conducted using existing tools?
- How do you mark cases and case elements which are relevant and important to facilitate further work?
- What additional information could computer tools usefully give you that they currently do not?

The relevance and importance of mooting

- Did you participate in a mooting society, mooting events or mooting competitions when you were training to be a lawyer?
- To what extent do you think that the process of preparing for a moot is similar to the process of preparing for a real legal case?

- What are the main ways in which mootng improves your understanding of the law and the legal process?

Contextual information

- To what extent does the use of computer-based platforms enhance the ability of a lawyer to understand and explore the implications of legal cases?
- Do you think that the visualisation of search results is an area where change is a priority at the moment?
- To what extent do legal search tools provide better access to relevant information than standard search platforms like Google or Bing?
- Do you make use of free legal tools and databases such as the British and Irish Legal Information Institute's online search facility and database?
- How do freely available legal information sources fit in with the specialist tools which you can use?
- In what ways, if any, are free tools less effective than the various commercial offerings that are available?

A.11 Lawyer survey questions

- What is the primary nature of your work in the law?
- If your answer was Other to the previous question, please briefly specify your job title or job description.
- What percentage of your working time do you spend preparing for litigation?
- Briefly, what are the top 5 activities that you are involved in when preparing for litigation?
- How often do you have to collaborate with other people in preparing for litigation?
- What are the top 5 activities in preparing for litigation which involve collaboration with other people?
- What online and offline tools and techniques do you employ to facilitate collaboration with others in your work?
- Do you use any of the following tools and techniques in order to facilitate collaboration?
- How important is effective collaboration in legal case preparation?
- How well do existing tools and processes which are available to you enable effective collaboration between the stakeholders in legal case preparation?
- What are the main challenges that you encounter when working with others to prepare a legal case?
- Do you typically use pen and paper or a digital device and software to take notes as you work?
- How much time do you spend taking notes whilst preparing a legal case?

A.12 Work role data from solicitor interviews

Roles and responsibilities
Builder
High-level classification for roles which involve work product output.
Argument builder
Writes argument skeletons and final legal brief to be put forward in court.
Note builder
Takes notes using pen and paper of important information during the preparation process. Brings disparate notes together into coherent information.
Flow builder
Prepares fact and transaction diagrams to make sense of the important facts in a case.
Researcher
High-level classification for roles which involve researching an issue in law.
General researcher
Consults stored output from general purpose tools like Wikipedia and Google to create a treatment of important information.
Specialist researcher
Uses stored output from specialist tools like HeinOnline and Westlaw to create a treatment of important information.
Searcher
High-level classification for roles which involve searching for and storing important information.
General searcher
Uses general purpose information tools like Wikipedia and Google to find and then store case reports, legislation and other legal documents for future research.
Specialist searcher
Uses specialist information tools like Lexis Library and Westlaw to find and then store case reports, legislation and other legal documents for future research.
Counsel

High-level classification for the only mandated split of responsibility between the members of a legal team.
Senior counsel
Responsible for either starting the submission to the court or finishing it, depending upon jurisdiction. Can be responsible also for the rebuttal if applicable.
Junior counsel
Responsible for either starting the submission to the court or finishing it, depending upon jurisdiction. Can be responsible also for the rebuttal if applicable.
Speech writer
High level classification with the responsibility of bringing information and arguments together into a coherent speech for delivery in front of the judge.
Project manager
High-level classification for roles which involve administrating a legal case, keeping the client file updated and billing clients.
File manager
Responsible for updating the file for a case with correspondence, billing information and other collateral for the purposes of auditing time and accounting for work to the client.
Communications manager
Responsible for liaising with outside experts, witnesses, the client and the opposition. Work here is heavily dependent upon email, written letters and telephone calls.

Table A1: Work roles data extrapolated from discussion in the solicitor interviews.

A.13 Drivers of collaboration from lawyer survey

Response 1	Response 2	Response 3	Response 4	Response 5
Discussing with other lawyers	Discussing with client	Discussing with others, e.g. witnesses, counterparties		
Conferring with colleagues for their opinion on reasonableness, prospects	Obtaining expert opinion eg on pension loss			
Conference with counsel	Comms with client	Telephone communication	Internal briefing	
Asking support staff to electronically page number the bundle of productions	Trainee assistance in preparing the bundle and making additional copies of the bundle	Discussing tactics with colleagues	Seeking advice from other colleagues in other specialisms (e.g.data protection) to ascertain the legal position	
Strategy	Law	Dealing with client	Witness statements	Mediation
Emails	Talking on phone	Meeting	Drafting, revising documents	Appearing in court

A. SETTING THE SCENE: CONTEXTUAL INQUIRY, INTERVIEWS AND SURVEY

Providing advice	Instructing counsel	Consult with counsel	Meetings with clients	Appearances in Court
Meeting and speaking with witnesses	Liaising with client	Liaising with counsel (if applicable)	Liaising with the other party	Liaising with the Employment Tribunal
Speaking, emailing with client	Speaking, emailing, meeting witnesses normally involved colleagues	Assistance from other departments, colleagues		
Productions	Witness lists, statements	Settlement negotiations		
Receiving instructions				
Witness statements	Productions	Court procedure	Discussing, settling case	Not sure
Taking instructions from the client	Identify witnesses	Obtaining documents and other evidence	Comms with the other party's solicitor	Comms with the court or tribunal
Evaluation with clients	Evaluation with litigation	Evaluation with expert witnesses, partners and advocates	Negotiation with third party's lawyers	Preparing and revising court papers
Preparing productions				

A.13. Drivers of collaboration from lawyer survey

Conferring with other team members (legal)	Working with the client to gather evidence	Production of expert evidence, reports	Working with counsel	Working with legal opponent to agree procedure, produce bundles of evidence
Letters and emails	Telephone calls	Meetings	Agreeing procedure	Agreeing content of joint documents
Document bundle production	Obtaining expert medical reports	Settlement discussions		
Discussing approach to litigation	Reviewing court papers	Attending meetings with colleagues, client, counsel	Attending court	Updating papers
Discussing legal position	Disputing legal position with other side	Preparing documentation		
Client	Barrister	Witnesses	Law Clerk	Court Office
Preparing Court Documents	Discussing Court Documents	Discussing Court Procedure	Dealing with Court Staff	Feeing
Conferring	Comms with the other side	Comms with the client		

A. SETTING THE SCENE: CONTEXTUAL INQUIRY, INTERVIEWS AND SURVEY

Sending emails	Telephone calls	Face to face meetings	Telephone conferences	Video conferencing
Agreeing collation of statements	Agreeing collation of documents	Reviewing pleadings	Preparing lines of questioning	Producing bundle of authorities



APPENDIX B

LARC - THE LEGAL
RESEARCH AND
COLLABORATION
PLATFORM

B.1 Final refined wireframe sketches for LARC

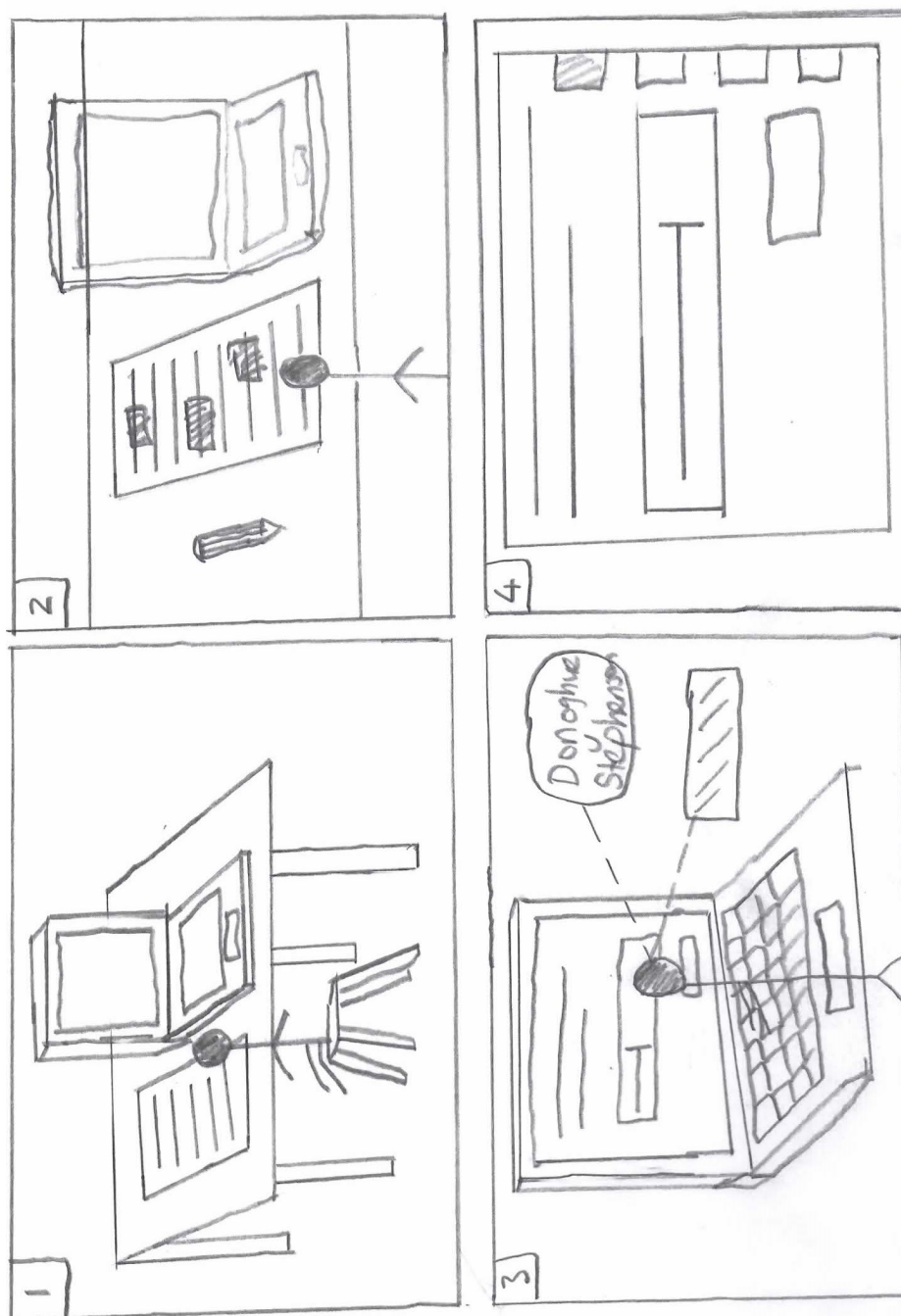


Figure B1: LARC Interface: Wireframes (set 1)

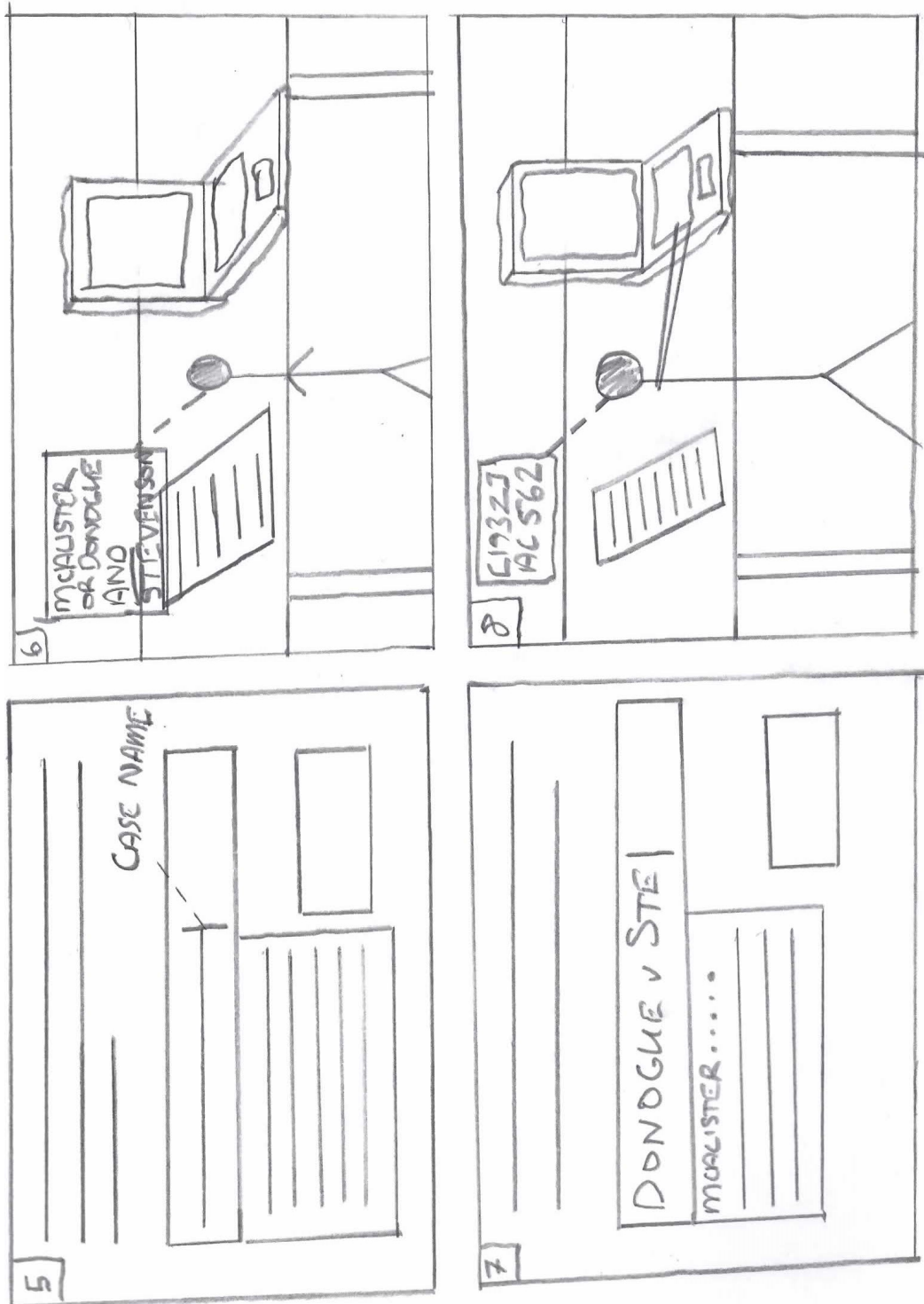


Figure B2: LARC Interface: Wireframes (set 2)



Figure B3: LARC Interface: Wireframes (set 3)



Figure B4: LARC Interface: Wireframes (set 4)

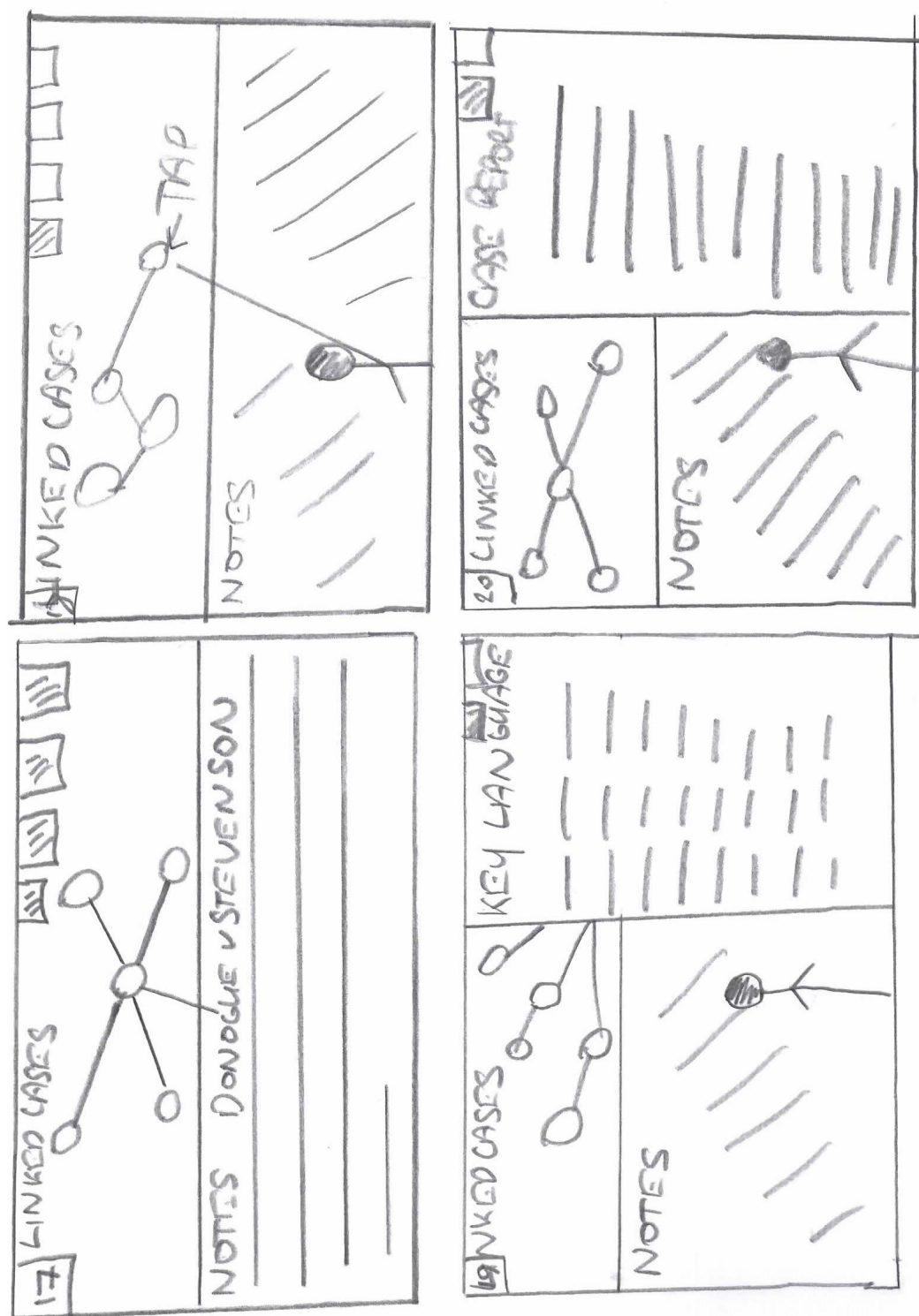


Figure B5: LARC Interface: Wireframes (set 5)

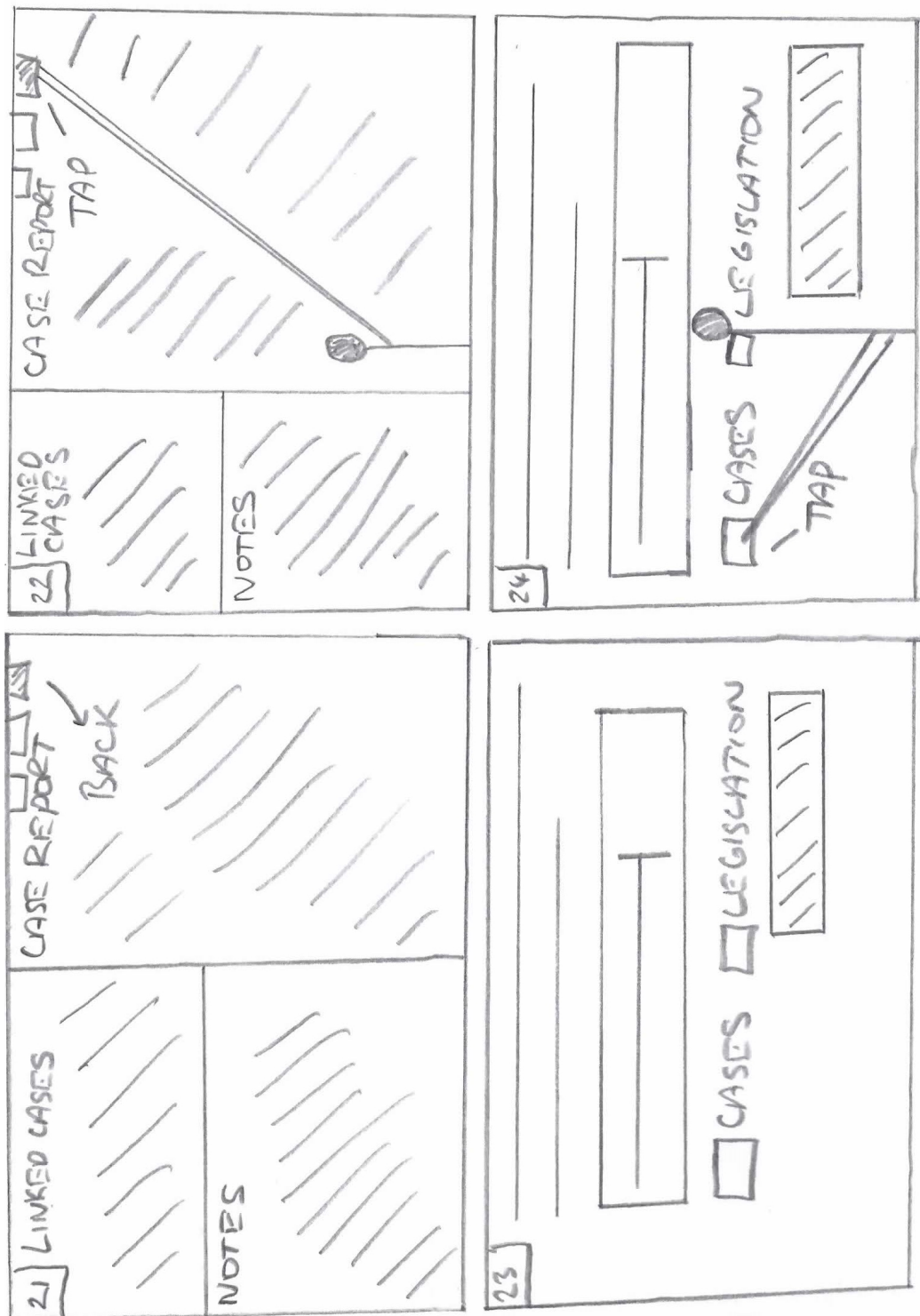


Figure B6: LARC Interface: Wireframes (set 6)

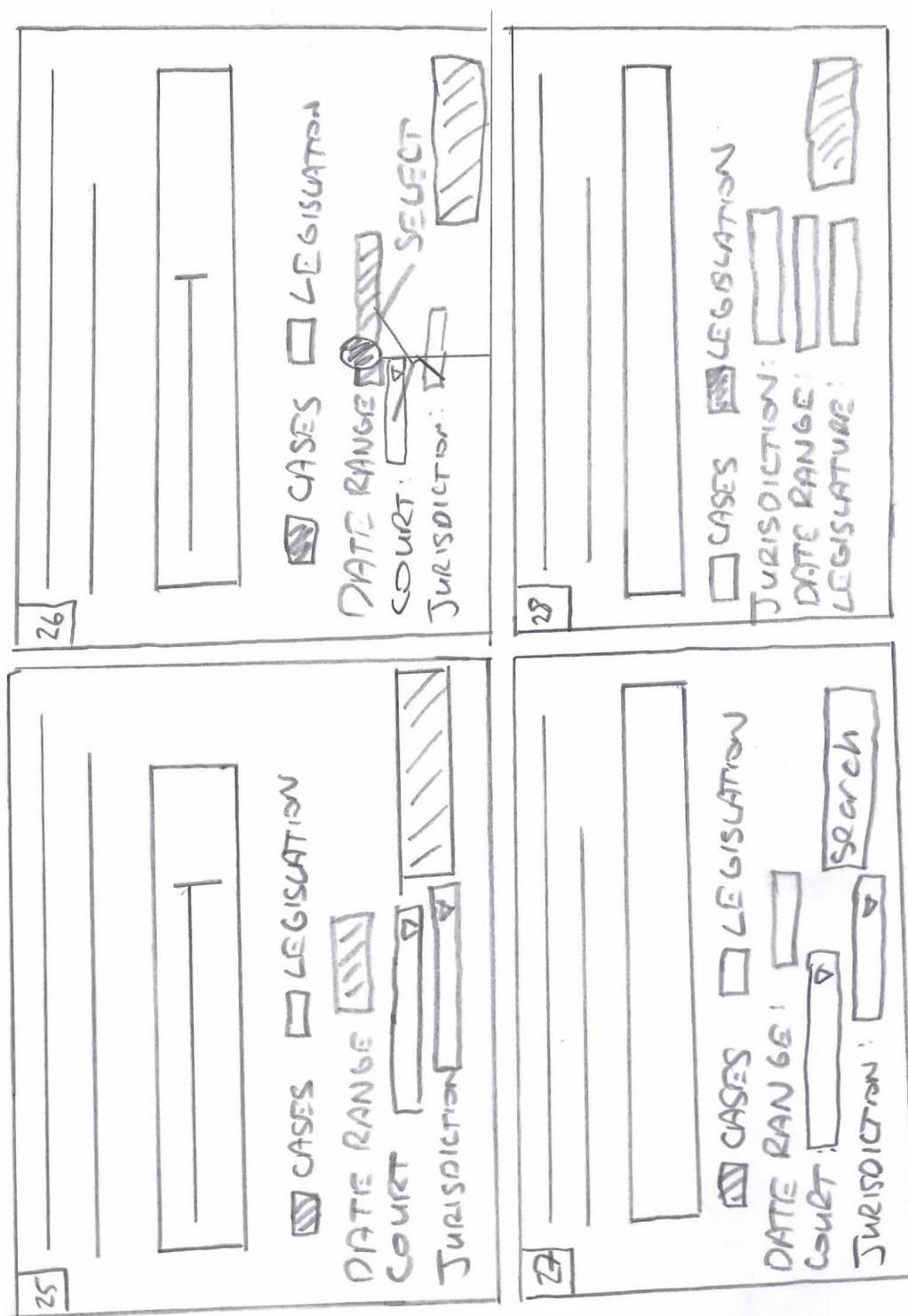


Figure B7: LARC Interface: Wireframes (set 7)

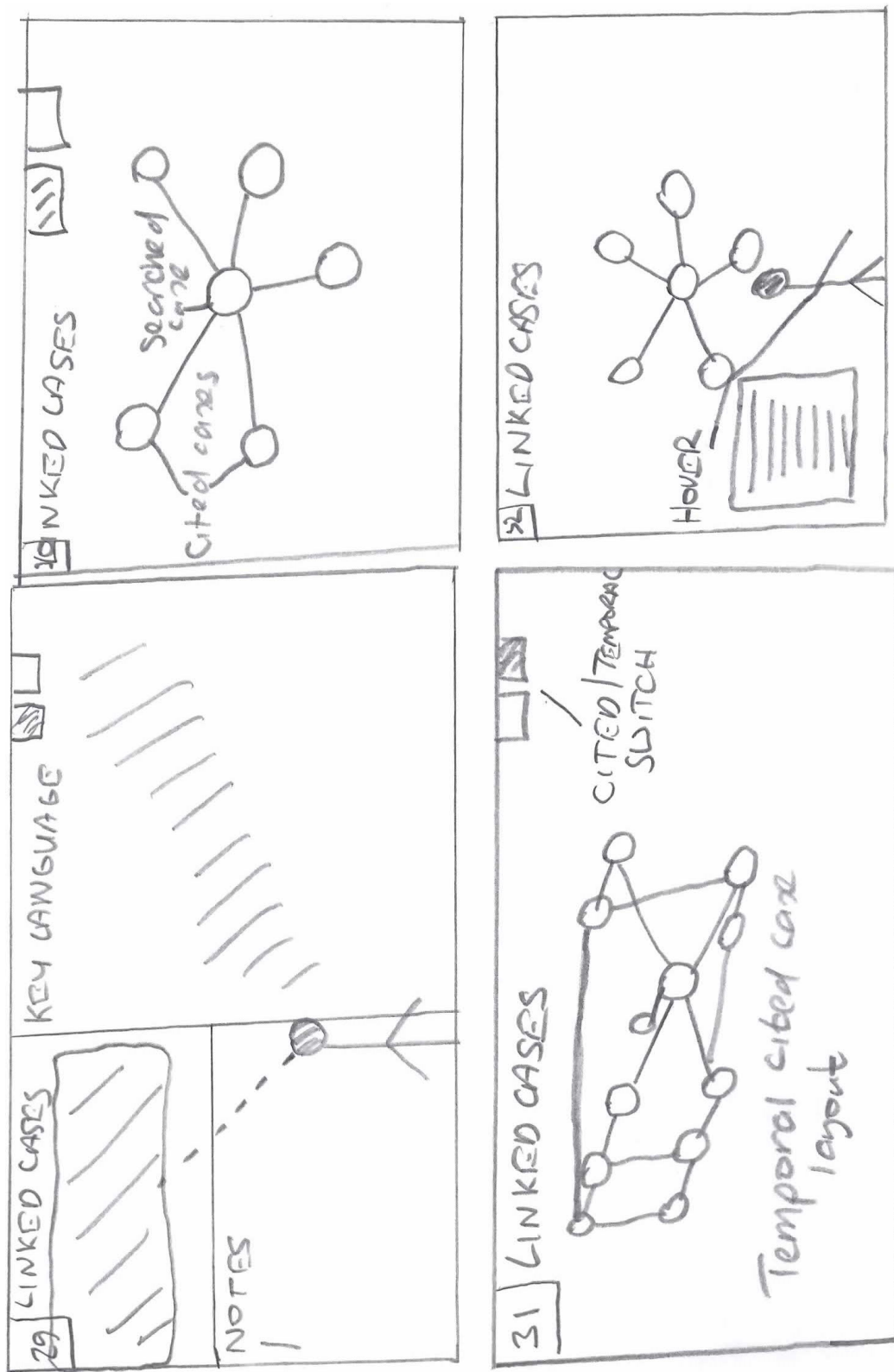


Figure B8: LARC Interface: Wireframes (set 8)

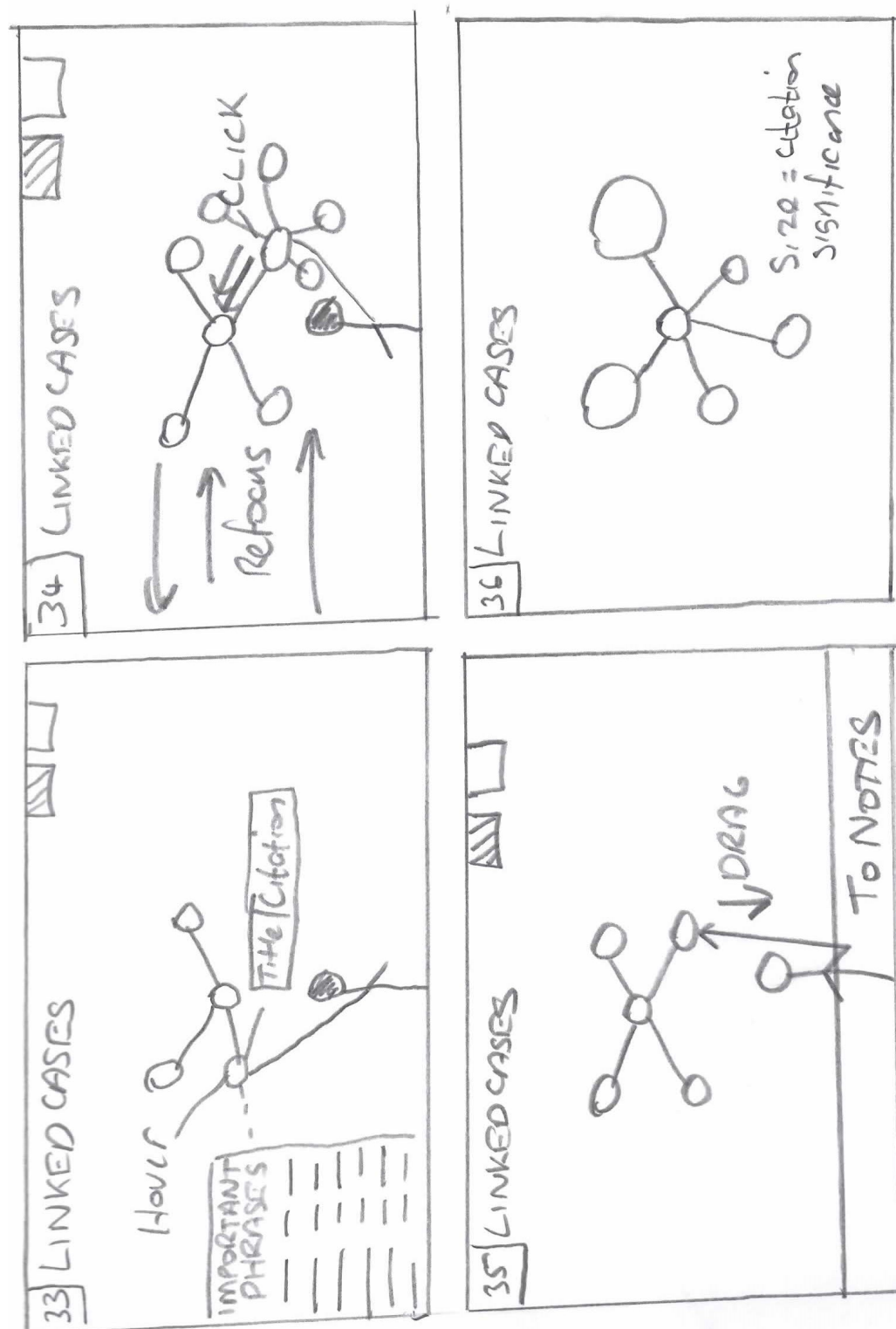


Figure B9: LARC Interface: Wireframes (set 9)

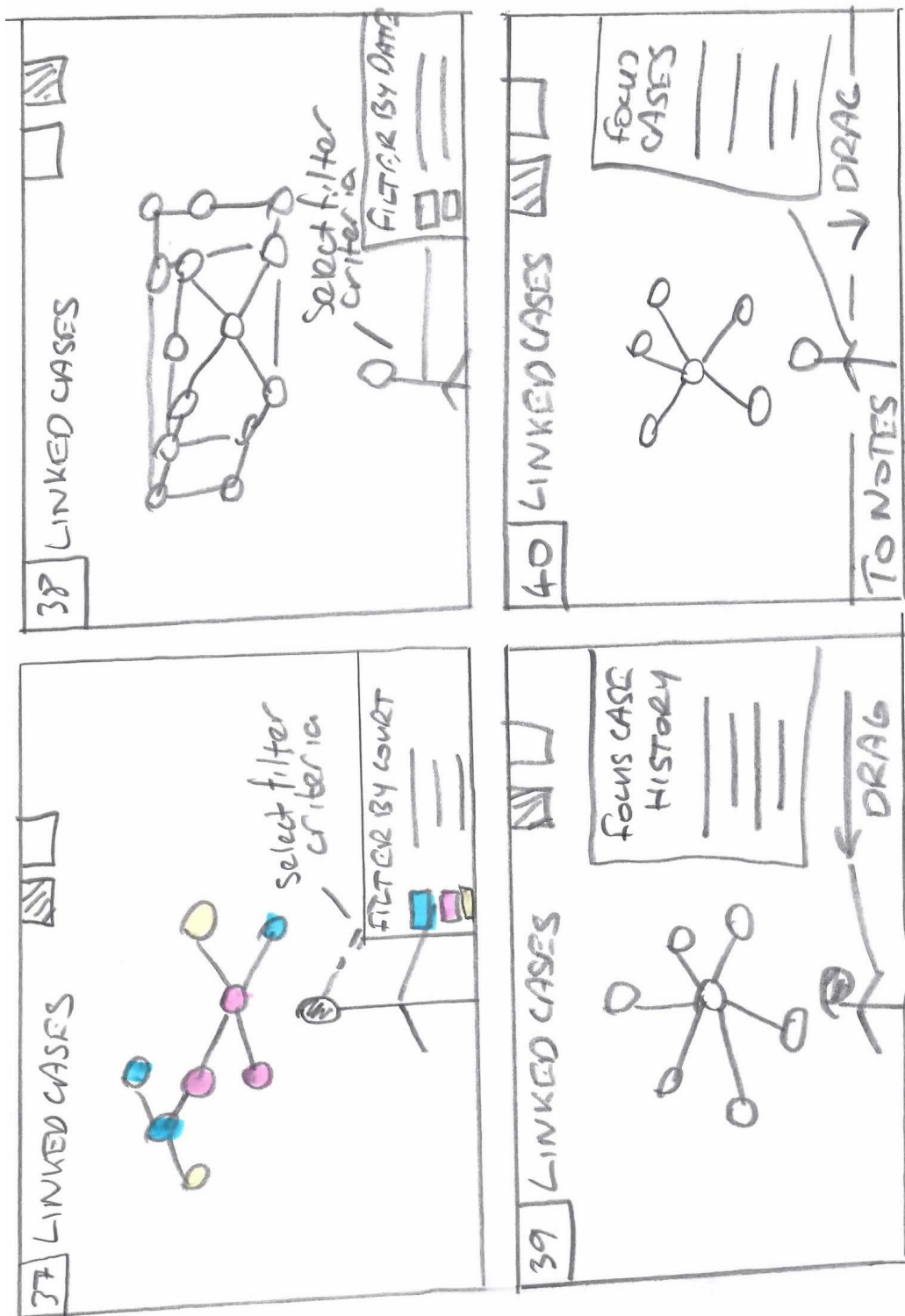


Figure B10: LARC Interface: Wireframes (set 10)

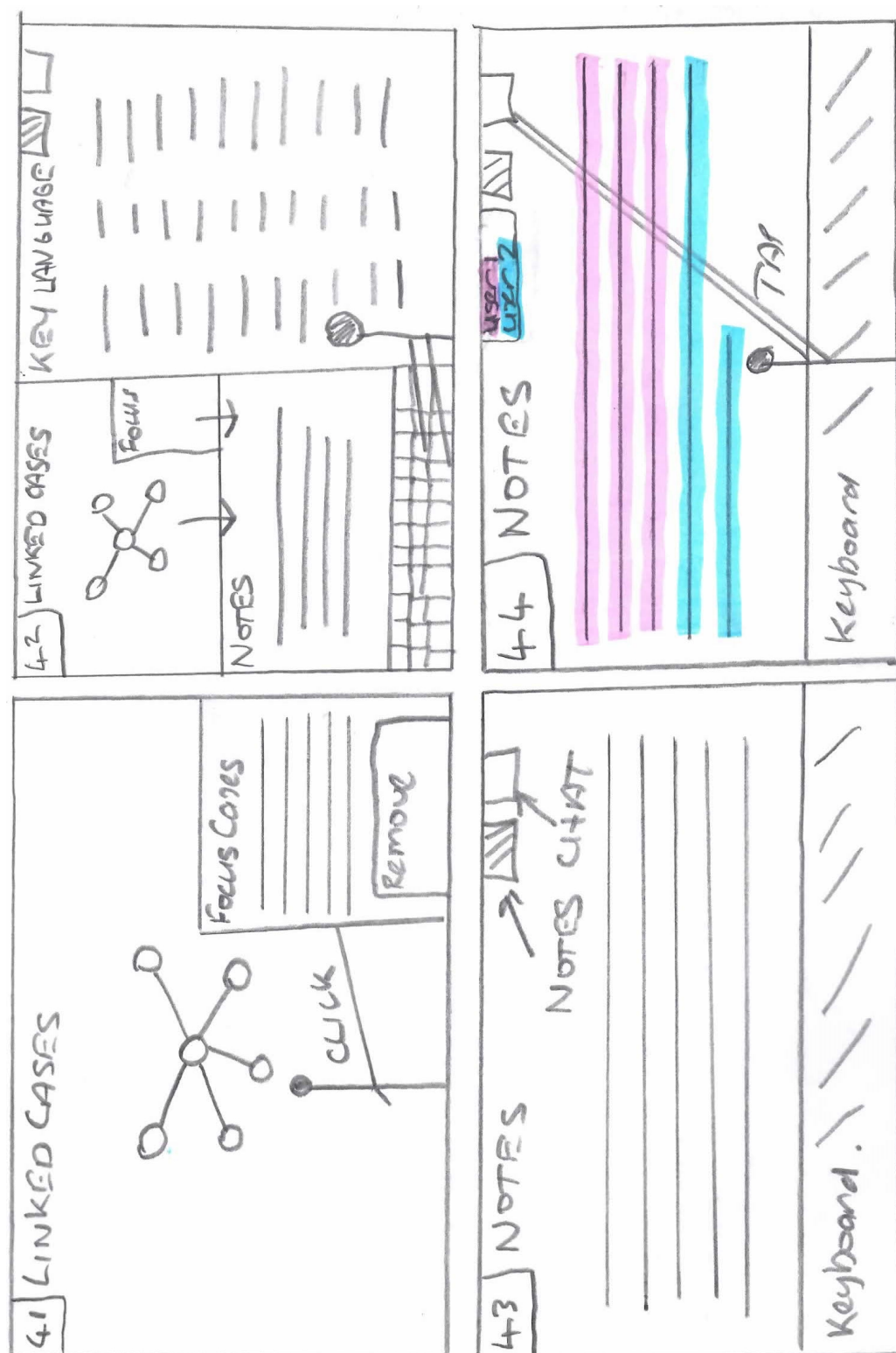


Figure B11: LARC Interface: Wireframes (set 11)

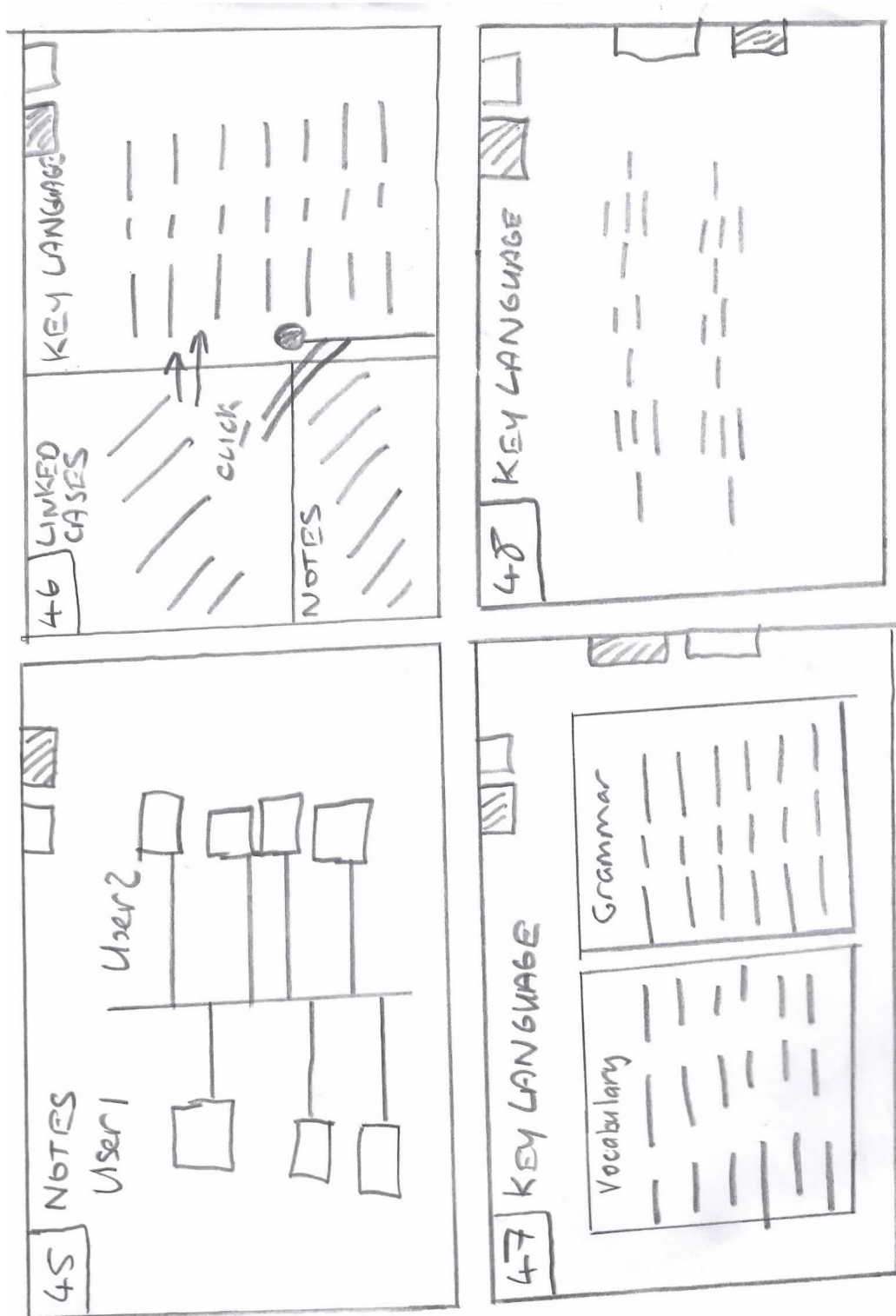


Figure B12: LARC Interface: Wireframes (set 12)

B.2 Wireframe details for LARC interface elements

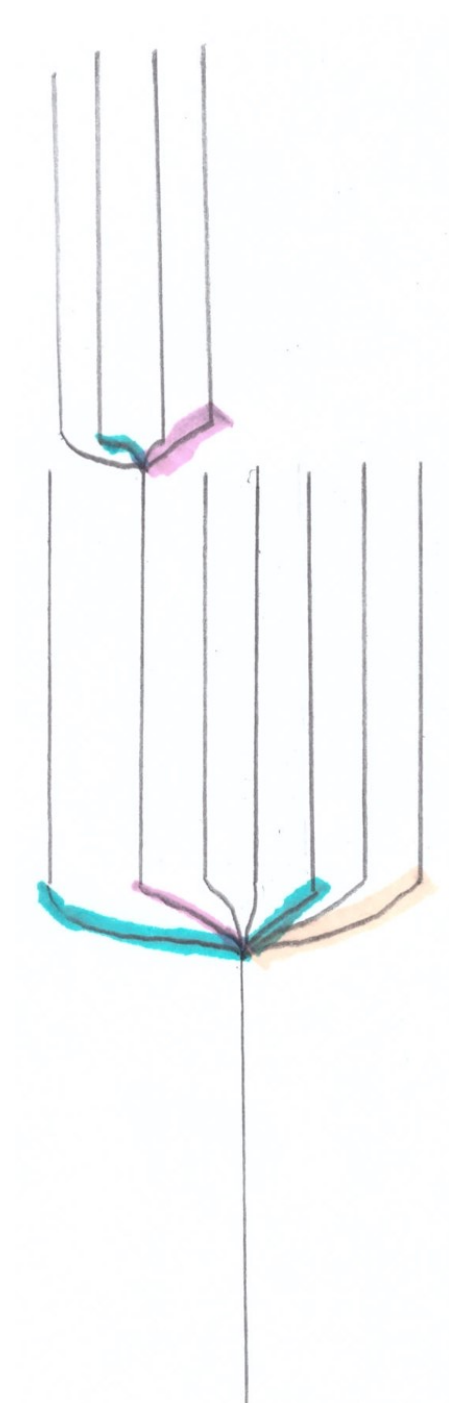


Figure B13: LARC Interface: Citation layout wireframe

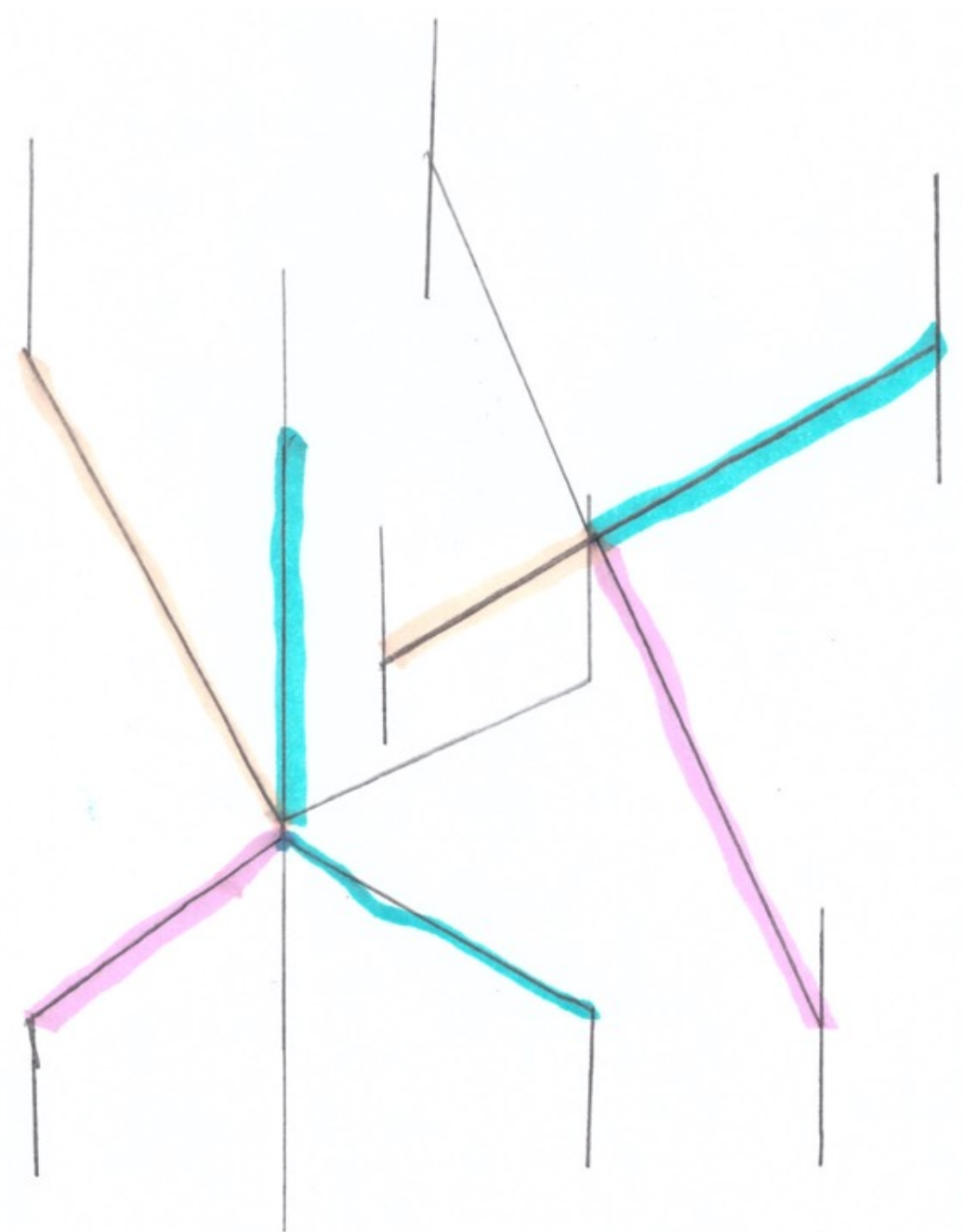


Figure B14: LARC Interface: On demand citation node loading

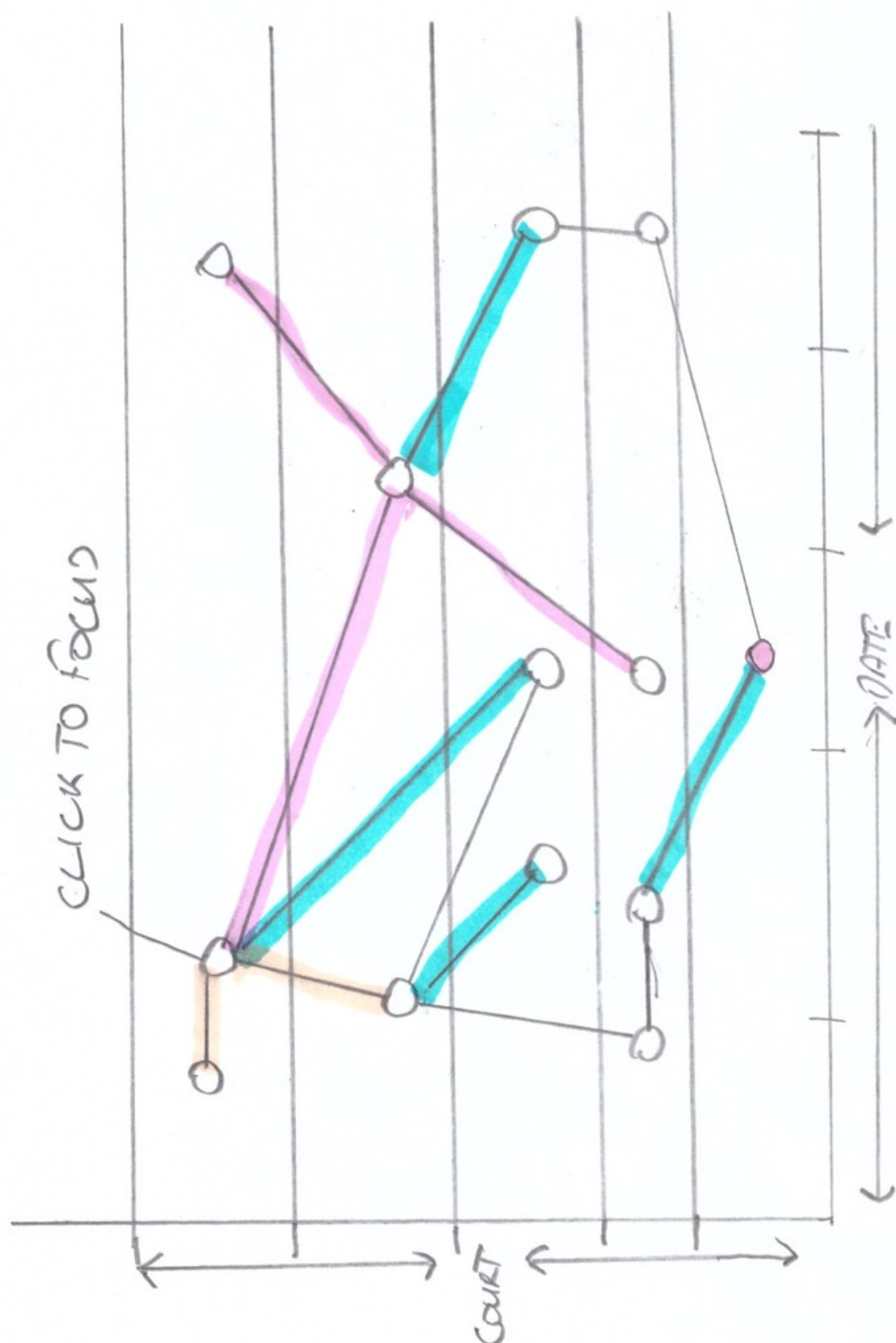


Figure B15: LARC Interface: Potential substrate wireframe

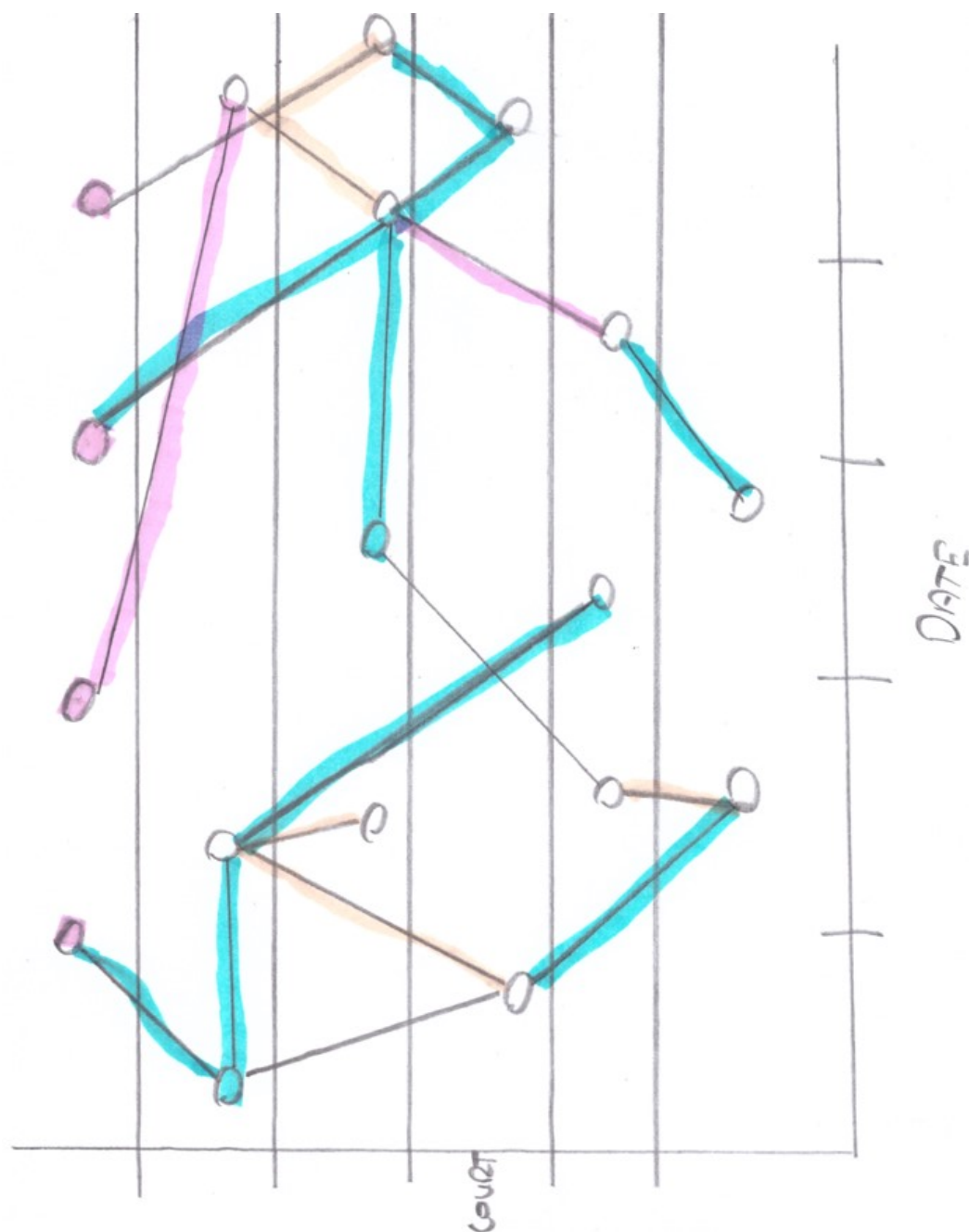


Figure B16: LARC Interface: Substrate elaboration wireframe



APPENDIX C

EVALUATION PRODUCTS

C.1 The evaluation questionnaire

Project Title

Collaboration Tools for Lawyers: LARC Evaluation

Q1 - What is the primary nature of your work in the law?

- I am a solicitor
- I am a barrister
- I am a partner in a legal practice
- I am an associate
- I am a paralegal or member of legal support staff
- I am a secretary or administrator
- I am an academic (lecturer/Professor etc)
- I am a law student
- I am a research fellow/postgraduate member of staff
- Other

Q2 - If your answer or one of your answers was Other to the previous question, please briefly specify your job title or job description.

Q3 - Problem Question

One day, while walking home, William trips and falls, damaging his knee. Several days later, while driving to work, he sees Victor crossing the road and brakes to avoid running into him. Unfortunately, due to the pain in his knee, he cannot fully press his brake pedal and, as a result, he runs into Victor. The collision occurs at a fairly slow speed and a normal person would only have suffered bruising as a result, but Victor has brittle bones and suffers two broken legs and a number of broken ribs. He is taken to the local hospital where, due to an administrative mistake, his right arm is amputated.

Advise Victor.

Instructions

- Log into LARC at <https://www.larclegal.com/>
- Use the username and password for the LARC system that you have received by previous email.
- Find an initial seed case on the system and create a new document for your work.
- Use LARC freely to research the problem question.
- Place relevant case citations, statutes and language into the document that you have created.
- Write an outline answer to the question in the LARC document.
- Once you have finished answering the question, please complete the following questions in this survey.

Time requirement

We envisage that you should spend no longer than one hour in writing an answer to this question.

Answer structure

The following points need to be discussed:

- breach of duty of care
- the rules of causation and the egg-shell skull rule
- a break in the causal chain with a possible intervening act by the hospital in amputating the arm
- *res ipsa loquitur* in relation to the liability of the hospital

Problems and help

The LARC system is in beta at the moment. You might run into errors or problems in using the software for this task. If you encounter an error or you need help, please email Evan Brown at **edb4@st-andrews.ac.uk**.

There are introductory videos available that demonstrate use of the LARC system. It may help you to watch some of these before attempting the question. The videos are available at <https://intro.larclegal.com/>

Q4 -I think that I would like to use LARC frequently.

- 1 - Strongly disagree
- 2
- 3
- 4
- 5 - Strongly agree

Q5 - I found LARC unnecessarily complex.

- 1 - Strongly disagree
- 2
- 3
- 4
- 5 - Strongly agree

Q6 - I thought LARC was easy to use.

- 1 - Strongly disagree
- 2
- 3
- 4
- 5 - Strongly agree

Q7 - I think that I would need the support of a technical person to be able to use LARC.

- 1 - Strongly disagree
- 2
- 3
- 4
- 5 - Strongly agree

Q8 - I found the various functions in LARC were well integrated.

- 1 - Strongly disagree
- 2
- 3
- 4
- 5 - Strongly agree

Q9 - I thought that there was too much inconsistency in LARC.

- 1 - Strongly disagree
- 2
- 3
- 4

- 5 - Strongly agree

Q10 - I imagine that most people would learn to use LARC very quickly.

- 1 - Strongly disagree
- 2
- 3
- 4
- 5 - Strongly agree

Q11 - I found LARC very cumbersome to use.

- 1 - Strongly disagree
- 2
- 3
- 4
- 5 - Strongly agree

Q12 - I felt very confident using LARC.

- 1 - Strongly disagree
- 2
- 3
- 4
- 5 - Strongly agree

Q13 - I needed to learn a lot of things before I could get going with LARC.

- 1 - Strongly disagree
- 2

C. EVALUATION PRODUCTS

- 3
- 4
- 5 - Strongly agree

Q14 - What did you like most about the LARC software?

Q15 - What did you dislike most about the LARC software?

Q16 - How would you change LARC to make it more useful in your work?

Q17 - How useful is the language search facility in LARC? How would you change it to make it better?

Q18 - Which features of LARC would be most useful for law students and trainee lawyers?

Q19 - Please tell us how useful you think LARC would be in your work as it stands.

- 1 - Not useful in my work
- 2
- 3
- 4
- 5 - Very useful in my work.